The Dissertation Committee for Blake Charles Borgeson certifies that this is the approved version of the following dissertation:

All-by-all discovery of conserved protein complexes by deep proteome fractionation

Committee:

Edward Marcotte, Supervisor

Andrew Ellington

John Wallingford

Claus Wilke

Vishwanath Iyer

All-by-all discovery of conserved protein complexes by deep proteome fractionation

by

Blake Charles Borgeson, B.S.E.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2016

Dedicated to all those whose life and health we are too late to save. And to my parents, who instilled in me the belief that we can do anything we set our minds to.

Acknowledgments

I want to thank the numerous people in my life without whom my journey into biology would have been so much less valuable and enjoyable. First, Edward Marcotte has found a way to combine and embody so many of the most wonderful things about science, in particular the boundless curiosity and excitement that draw us to it in the first place. I could not have found a better place to discover how biology, math, and computation can play together to such wonderful ends. I thank my committee members for their understanding, support, and encouragement along my unusual trajectory.

I must heartily thank Cedric and Martin for teaching me and inspiring me in this new art of scientific computing and machine learning, and Taejoon, who greatly sped my understanding of mass spectrometry processing and computing on UT's enormous cluster. I thank my collaborators: Traver for setting me down the proper path, Andrew, Cuihong and Sadhna for their massive scientific efforts and countless hours on Skype, Pierre for digging up so many mountains of data, Ophelia, Alice, Anna, and Fan for their enthusiasm to contribute to and study protein complexes. I thank Jeremy for trying admirably to teach me the necessary skills for benchwork, Dan and Andrew for explaining the intricacies of proteomics, and Kevin for his beautiful analyses and for forcing me to check in and explain my code. I thank Peggy for getting us scientists to get out and socialize, and Gabe for hosting so

many excellent parties at my house and helping me out in a pinch so many times. And I thank Jag, Alex, Angela and Aashiq for boosting the level of intellectual stimulation in the lab.

Finally, I can't help but express some of the deep gratitude for the many relationships that have kept me connected to the fabric of humanity I love. I thank Jacob and Jimmy and Craig for enduring friendships that help me remember how great and fun the world is. I thank Dan and JR for their support and interest in my pursuit of understanding biology, and Chris, who has tolerated, appreciated, even encouraged the divided attention it required to see this chapter of my career through to the end. I thank those whose relationship with me suffered on account of my work and ambitions. And of course, I thank my parents for their endless enthusiasm for my life and work.

All-by-all discovery of conserved protein complexes by deep proteome fractionation

Publication No. _____

Blake Charles Borgeson, Ph.D.
The University of Texas at Austin, 2016

Supervisor: Edward Marcotte

Stable assemblies of proteins, known as protein complexes, execute a large fraction of cellular processes required to sustain life. A functional and mechanistic understanding of these assemblies will provide a more comprehensive understanding of an organisms genes and elucidate a more complete picture of cellular processes, particularly those involved in development, aging and disease. While recent progress has mapped protein complexes in budding yeast and some bacteria, efforts in animals are restricted to subsets of the proteome, leaving most animal protein complexes undetermined. Co-fractionation offers compelling efficiency gains in identifying pairwise protein interactions and complexes, but it requires significant computational efforts to fully exploit. In this work, I describe the computational methods and infrastructure I developed to identify conserved protein interactions and complexes from a massive set of mass spectrometry data from nine species and the computational and biological analysis I performed with my collaborators.

These efforts include building a mostly automated pipeline to process and integrate large quantities of mass spectrometry data from multiple species and developing improved methods to predict co-complex interactions and cluster them into complexes. The conserved animal complex map produced using this pipeline and methodology has already yielded dividends in supporting biological discoveries. Scaling the approach more broadly will enable rapid mapping of the previously uncharted interactomes in any chosen species.

Table of Contents

List of Tables

List of Figures

Chapter 1

Background: protein interactions, complexes and networks

1.1 Genes to Genomes

It is striking to recall that publication of the first eukaryotic genome, the yeast *S. cerevisiae*, occurred only twenty years ago (Goffeau et al., 1996). At only twelve million base pairs, the yeast genome is dramatically smaller than the human genome at three billion base pairs. Yet, publication of the initial draft (>90% complete) of the human genome followed only five years later (HGP, 2001). The genomic revolution in the ensuing years, driven in part by dramatically decreasing sequencing costs, yielded the publication of a growing set of organisms, from mice to flies to worms, with the list now including hundreds of organisms in the animal kingdom alone.

Molecular biology and genetics were thriving fields of research prior to the emergence of fully-sequenced genomes. However, full knowledge of the genomes of many organisms was a boon for research in molecular and cellular biology. By providing a complete ingredients list, this was a critical step towards unravelling the underlying substrates for biological processes. Full genomes provided biologists with a starting point for many additional questions, from investigating and

characterizing previously unknown genes, to querying the function of suspected particular regulatory regions, to unmasking the consistency and disparity of patterns observed across promoters, splice sites, untranslated regions, and protein domains (ENCODE, 2012). Genomes also served as the basis to advance evolutionary studies, with even a measure as simple as sequence conservation driving substantial progress in the field of phylogenetics and with greater understanding of both the evolutionary relationships among species and the mechanisms of evolutionary processes themselves (Nakhleh, Ringe, and Warnow, 2005; Lynch et al., 2007). Further, creative methods to use genomes across organisms has led to broad insights about genes' functions, for instance exploiting the observation that genes that evolve together are more likely to share function[1] (Tillier and Charlebois, 2009). In numerous ways, research in evolution and molecular and cellular biology has benefited enormously from the increasing completeness of genomes across species.

1.2 Genomes to interactomes

While fully sequenced genomes are a critical step towards understanding, this list of ingredients across organisms is an oversimplification analogous to early models of cells as bags of proteins and other chemicals. The field has long understood that multiple forms of coordination exist at many levels between genes and cellular organelles, as a significant organizing principle in cells is protein complexes. Research into the molecular mechanisms of cellular behavior in humans

[1] http://science.sciencemag.org/content/285/5428/751.short

and other organisms has provided dramatic evidence for the importance of protein complexes to a vast number of cellular processes (perspectives in Alberts, 1998; Hartwell et al., 1999). Protein complexes represent a layer of organization that connects individual genes and proteins to their mechanistic cellular functions and to the phenotypic effects to which they contribute. This modular understanding of molecular biology is key to advance our understanding of the organization of cellular processes and the link from genotype to phenotype, the evolution and conservation of biological systems, and how this manifests in human disease.

1.2.1 Incompleteness of protein interactions and complexes

While the relevance of protein complexes in human biology has long been appreciated, knowledge of protein complexes remains far from complete across nearly all of biology, including in humans, similar to the state of genomes before the advent of the first fully-sequenced genomes near 2000 (Adams et al., 2000). Historically, the discovery of protein complexes has mostly been limited to isolated studies by various researchers investigating particular cellular processes or diseases of interest. Even with the incomplete and unreliable state of current knowledge of protein complexes, studies have already yielded insights into cellular organization, gene function, and evolutionary mechanisms.

1.2.2 Complexes and functional organization

As early as 2001, the structure of protein interaction networks have been investigated as a source of information on overall cellular organization. One of the

early observations of this nature was the observation of a strong correlation between the degree, or number of interacting partners, of a protein in an interaction network and its essentiality to cell survival (Jeong et al., 2001). As an organizing feature in biological systems linking multiple genes to their specific activities and functions, proteins participating in a complex together tend to share functions. Multiple studies in yeast revealed that protein complexes share characteristics such as knockout phenotypes and essentiality (Gavin et al., 2006; G Traver Hart, Lee, and Edward M Marcotte, 2007). This sharing of function can be applied in reverse to assign function to genes whose function is unknown or poorly annotated. This principle of guilt-by-association, which in this context infers function based on the fact that two interacting proteins are more likely to share function than non-interacting proteins, has been applied for over a decade and has continued to provide novel biological insights (Oliver, 2000; Wang and Edward M Marcotte, 2010). Recently, researchers mapped protein complexes in the bacteria *E. coli* and used guilt-by-association to propose functions for nearly one-third of its genome whose functions were at the time undocumented (Hu et al., 2009). The study of protein complexes also has a history in improving our understanding of human diseases. Relying again on the principle of guilt-by-association, researchers used inferred human protein complexes and existing disease-associated genes in a machine learning framework to predict thousands of additional links between genes and diseases (Lage et al., 2007). Other groups have employed similar strategies using existing protein-protein interaction datasets (e.g., Fraser and Plotkin, 2007). Still others have utilized protein complexes as a resource to biologists performing high-throughput functional inter-

action studies as a means to interpret results and provide additional statistical power to identify weak signals (Vinayagam et al., 2013).

1.2.3 Complexes and evolution

Protein complexes can also reveal novel relationships between genes and pathways across evolution. As soon as substantial numbers of protein interactions started to accumulate in yeast and some commonly-studied bacteria, computational work exploited these interactomes to infer conserved pathways across these distant species (Kelley et al., 2003). This approach was extended to interactomes of flies and worms, as significant numbers of interactions were identified in those species (Sharan et al., 2005). Later, independent groups used alignment of protein interactions across evolution to infer functional orthologs—genes serving the same function across species—in cases where sequence-based methods had not yet identified these genes as orthologs (S Bandyopadhyay, 2006; Singh, Xu, and Berger, 2008). Similar to the discoveries uncovered with the advent of full genomes, more complete maps of protein complexes spanning evolution will stimulate many new organizing principles.

1.3 History of mapping protein interactions

Convinced of their value as a lens to understand biological structure and function, scientists have for decades sought to identify protein-protein interactions and complexes, as well as develop new experimental techniques and computational tools to aid them in their search. Early knowledge of protein complexes emerged

slowly through low-throughput experimental studies isolated to single complexes, mainly employing biochemical separation methods, such as gel filtration, affinity chromatography and affinity electrophoresis, to validate prior hypotheses rather than in an exploratory fashion (Beeckmans, 1999).

1.3.1 Yeast two-hybrid

Beginning two decades ago, groups began using a systematic approach called yeast two-hybrid (Y2H) to screen for direct physical binding between two proteins by attaching one protein called the "bait" to the DNA-binding domain of a transcription factor for a marker gene, and the other protein known as the "prey" to the activation domain of the same transcription factor, thus activating transcription of the marker gene only when the two proteins bind one another (S Fields and Song, 1989; Brent and Ptashne, 1989). Two groups applied this approach to a subset of several thousand proteins of the human proteome (Stelzl et al., 2005; Rual et al., 2005). Another group applied this approach to query pairwise interactions between approximately 14,000 human proteins (Rolland et al., 2014). Compared to other approaches, Y2H experiments have the advantage of being highly unbiased in the interactions they investigate, have equivalent effectiveness regardless of a protein's native expression level, and are immune to difficulties other approaches have in identifying interactions involving membrane-bound proteins (Typas and Sourjik, 2015). Use of Y2H experiments to map protein complexes, however, suffers from several challenges. First, as only direct binary physical interactions are measured, complexes themselves are never isolated. Isolating them from binary interactions is

not straightforward, so they are often not attempted despite the proteome-wide scale of interactions measured (Rolland et al., 2014). Second, Y2H experiments are not specific to the native environment and the context in which the interaction usually occurs, including non-physiological expression levels and a complete lack of the cellular environment, and thus increasing the potential to capture a large number of false positivesproteins who stick together in a Y2H experiment but would never interact physiologically due to different expression levels, timing, tissue specificity, or subcellular localization of proteins. Third, such experiments require expressing hybrid versions of the proteins of interest, which raises the question of whether the proteins are altered in a way that affects their native interaction tendencies.

Beyond the above obstacles to using Y2H experiments specifically to identify protein complexes, Y2H experiments may also fail to replicate. Post-publication analysis of published putative protein-protein interactions identified in the same organism by different labs using Y2H revealed surprisingly low overlap (Von Mering et al., 2002; Güldener et al., 2006). Recent efforts now include multiple rounds of screens and multiple stages of validation of observed interactions, which may reduce their false-positive rate, though at the cost of increasing experimental efforts (Rolland et al., 2014). Leaders in Y2H screening remain optimistic that the method will yield a fully mapped human interactome (Vidal and Stanley Fields, 2014).

1.3.2 Pull-down and mass spectrometry

Several other groups have undertaken a different approach to overcome these challenges inherent with Y2H screening using pull-down experiments em-

ploying an affinity purification step followed by mass spectrometry (AP-MS), which was first successfully performed at a broad scale in yeast (Gavin et al., 2006; Krogan et al., 2006). In this approach, a tagged version of the protein is pulled from solution along with any other stably bound proteins, and then this mixture is isolated and the other bound proteins in the putative complex identified using mass spectrometry (Figure 1.1). While overcoming many of the obstacles presented by

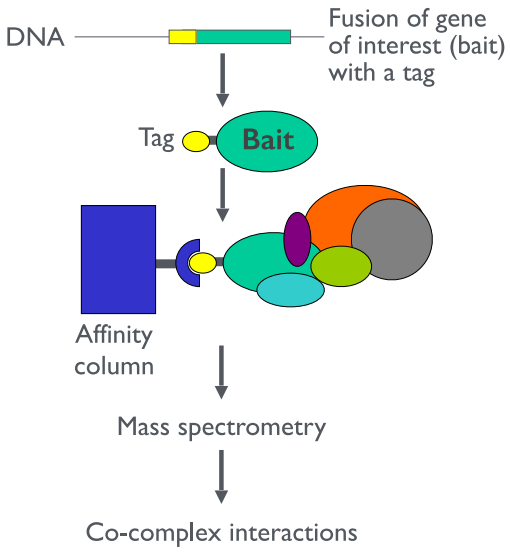


Figure 1.1: **Affinity purification coupled to mass spectrometry (AP-MS).** This powerful method of identifying co-complex interactions requires tagging or acquiring and validating antibodies for each gene of interest.

Y2H in mapping complexes, AP-MS, is labor-intensive by requiring expression of a tagged version of the protein, and again fails to observe protein interactions in its fully-native state. Despite these difficulties, the approach's success in yeast has encouraged multiple groups to apply it to mapping complexes in targeted subsets

of the proteome in a few species including fly and human (Guruharsha et al., 2011; Malovannaya et al., 2011). Coincident with this work, several groups have published larger attempts at mapping human protein complexes through AP-MS, performing pulldowns of between one and three thousand bait proteins and identifying thousands of co-complex interactions (Hein et al., 2015; Huttlin, Lily Ting, et al., 2015). One common feature of these studies is the massive scale and complexity, requiring expression of thousands of tagged proteins, and running and analyzing thousands of mass spectrometry experiments.

1.3.3 In silico prediction methods

In response to the immense effort, expense and complexity involved in existing experimental methods to identify protein interactions, a number of groups have taken in silico approaches to predicting interactions based on other biological data. Since 2003, one group integrated the limited existing co-complex interaction data, Y2H data, and other biological data indicative of interactions at the time in a machine learning framework to predict pairs of proteins likely to be contained within the same complex in yeast (Jansen et al., 2003). Similarly, an approach taken several years later drew upon both genetic interactions and AP-MS data in yeast to increase confidence and coverage beyond either dataset alone to identify protein complexes (Sourav Bandyopadhyay et al., 2008). Although coverage of their predicted interactions was highly limited by the paucity of reliable input data at the time, their work showed the power of machine learning and promise of incorporating various sources of biological data in conjunction with signal-rich experimental data, a prin-

ciple I have sought to leverage in my own work. More recently, one group achieved high coverage in predicting yeast co-complex interactions, this time structurally, by making use of the reasonable coverage of known protein structures in yeast and extending these through homology modeling (Zhang et al., 2012).

1.3.4 Co-fractionation for massively parallel native protein complex discovery

Lacking the approaches described above as a means to identify protein complexes in their native cellular milieu, unbiased and at high coverage across species with limited structural knowledge of proteins and without the effort and concerns involved in tagging thousands of proteins or developing antibodies, we developed a novel approach to discover native protein complexes. The collaboration between the Marcotte and Emili labs recently led to the development of co-fractionation as a method for scalable discovery of native protein complexes. A pilot project, published while this work was underway, involved approximately a dozen fractionations of two human cell lines and revealed a significant number of human protein complexes (Havugimana et al., 2012). By developing approaches to incorporate data and target predictions from multiple species at a much larger scale, the work described in the following chapters identifies and explores novel interactions and complexes.

The critical computational infrastructure, tools and analysis that underlie this undertaking are the subject of this work, as detailed and described in the following chapters. Chapter 2 describes the proteomics data processing infrastructure

I built to support automated processing of raw data from thousands of mass spectrometry experiments using open source tools and search databases, running on a university-owned cluster infrastructure. Chapter 3 explores the methods I designed to identify co-complex protein interactions from fractionation experiments. In Chapter 4, I discuss the methods I developed to cluster protein interactions to produce a large and accurate map of conserved animal protein complexes. In Chapter 4, I also discuss the multiple insights and discoveries enabled directly by this map.

Chapter 2

An automated pipeline for consistent proteomics data processing across multiple species

This first major component of my work involved the development of a computational processing pipeline sufficient to derive protein abundances from mass spectrometry data on a nearly unprecedented scale. Biological sample preparation and mass spectrometry were led by Cuihong Wan at the University of Toronto. Additional sample preparation and mass spectrometry was carried out by Ophelia Papuolas in the Marcotte lab, with assistance from Dan Boutz, and raw mass spectrometry data from prior human cell samples was contributed by Pierre Havugimana. The research, development and analysis described in this chapter are my own work. While this chapter is original and not based on published work, portions of this work were described in a publication[1] by my collaborators in the Emili lab, on which I was a co-author.

[1] Wan, C. et al., 2012. ComplexQuant: High-throughput computational pipeline for the global quantitative analysis of endogenous soluble protein complexes using high resolution protein HPLC and precision label-free LC/MS/MS. Journal of Proteomics, pp.110. All work described in this chapter is my own.

2.1 Background

2.1.1 Mass spectrometry for protein identification and quantitation

Proteomics experiments are now feasible at a scale orders of magnitude beyond what was possible only a decade ago similar to the present situation in genomic research. The development of 2-D gel electrophoresis (O'Farrell, 1975) stimulated interest in the precise measurement of the state of the entire proteome. Tandem mass spectrometry was effectively applied to protein mixtures a decade later (Hunt et al., 1986), followed by the addition of liquid chromatography for greater sensitivity and a database lookup algorithm (Eng, McCormack, and Yates, 1994) to enable identification of hundreds of known proteins in a single run in the next decade (Link et al., 1999). These developments led to the general mass spectrometry experimental and data analysis workflow for proteomics still widely used today. Protein fragments known as precursor ions enter into the machine, generally through a liquid chromatography step to separate the mixture for higher sensitivity, for analysis and produce "MS1" mass per unit charge spectra. Selected precursor ions are moved into a gas collision chamber and then are broken into smaller peptide fragments, which are passed to a second analyzer producing "MS2" spectra. A proteome sequence database for the species under study is prepared and subjected to in-silico digestion and fragmentation according to process models. Spectra measured experimentally are compared against those computed from the database and yield scored peptide-spectral matches, which are filtered and aggregated according to the desired quality control protocols to produce the final protein identification and quantitation.

2.1.2 Processing mass spectrometry data

Since that time, the field has seen experiments of increasing scale and commensurately improved coverage of the genome, in terms the number of identified and quantified proteins. Recent efforts to map the human proteome by leading proteomics labs, coincident with our own work, have approached the scale of this project (Kim et al., 2014; Wilhelm et al., 2014). However, as of 2013, even the largest-scale attempt at capturing the human proteome only analyzed 127 experimental LC/MS-MS runs and identified around 400,000 peptide-spectral matches (Beck et al., 2011). These experiments are typically analyzed by dedicated proteomics experts in leading laboratories using one or a few high-powered workstations. Therefore, we developed an alternative approach that would shepherd our 6,387 LC/MS-MS runs through workstations and acquire more than 10,000,000 filtered peptide-spectral matches.

2.2 Approach

2.2.1 A proteomics project of massive scale and complexity

Proprietary data formats are standard for output from modern mass spectrometry machines requiring libraries and/or licenses installed on the computers used to process these raw files into protein quantitations. Microsoft Windows-based libraries are widely used for this aspect of the workflow, in particular using the popular software MaxQuant for protein quantitation (Cox and Mann, 2008). However, our team had several project requirements that made the standard workflow less feasible.

First, the sheer quantity of data acquired by our mass spectrometry experiments dwarfs the scale of data in our human-only pilot, as described further in Section 2.3, below. The size of our experimental multi-species dataset is rivaled only by a handful of contemporaneous proteome and interactome surveys, employing multiple sample conditions and multiple biochemical fractionation steps in order to achieve greater sensitivity to identify rarer proteins (Hein et al., 2015; Huttlin, Lily Ting, et al., 2015). Second, we wanted to capitalize on the multiple leading mass spectrometry database search platforms and leverage MSblender to probabilistically combine these search results for improved coverage and accuracy (Kwon et al., 2011). Our lab in collaboration with others previously showed that probabilistic integration of results from multiple search databases improves the sensitivity and accuracy of proteomic analyses. Given our goal to maximize coverage and quality across multiple species, we wanted to take advantage of this improvement. Finally, proteomics projects spanning multiple species are rare, and nothing at this scale had been attempted to date. An important boost to more effectively compare and integrate data across species is consistent protocols, both experimentally and computationally. Since protein sequence database preparation to quality control and processing choices are central to the aim of this project, consistent handling, protocols, and treatment across species was required.

2.2.2 Running on open source computational infrastructure

While leading labs that perform such proteome survey experiments often employ computational proteomics experts dedicated to develop their own protein

quantitation software and execution of processing on clusters of dedicated full-access Windows machines, I developed instead an approach to process thousands of raw mass spectrometry files in an automated manner using our existing available computational resources. The University of Texas has a large linux-based computing resources at its disposal in TACC[2]. So, I developed an automated processing pipeline that solely relied on open-source components. The resulting pipeline enabled conversion, processing, and scoring using thousands of cores in parallel on TACC (Figure 2.1).

2.2.3 Preparing protein sequence databases

Protein sequence database preparation was a high priority and essential to enable me to compare and integrate proteomics data across species. One decision with surprisingly large downstream effects on ease and efficiency for cross-species comparison was stipulating gene and protein identifiers and protein sequence sources. Since ENSEMBL has a gene-focused perspective on identifiers more conducive to cleaner orthology mappings and covers a significant portion of our species under study, I chose to rely on protein sequences and identifiers from ENSEMBL for the available five species and acquired other protein sequences and gene and protein identifiers of interest from relevant organism-specific databases (see Section 2.5).

To improve the coverage and quality of identified complexes and to support the discovery of interactions using co-fractionation experiments across many

[2] https://www.tacc.utexas.edu/

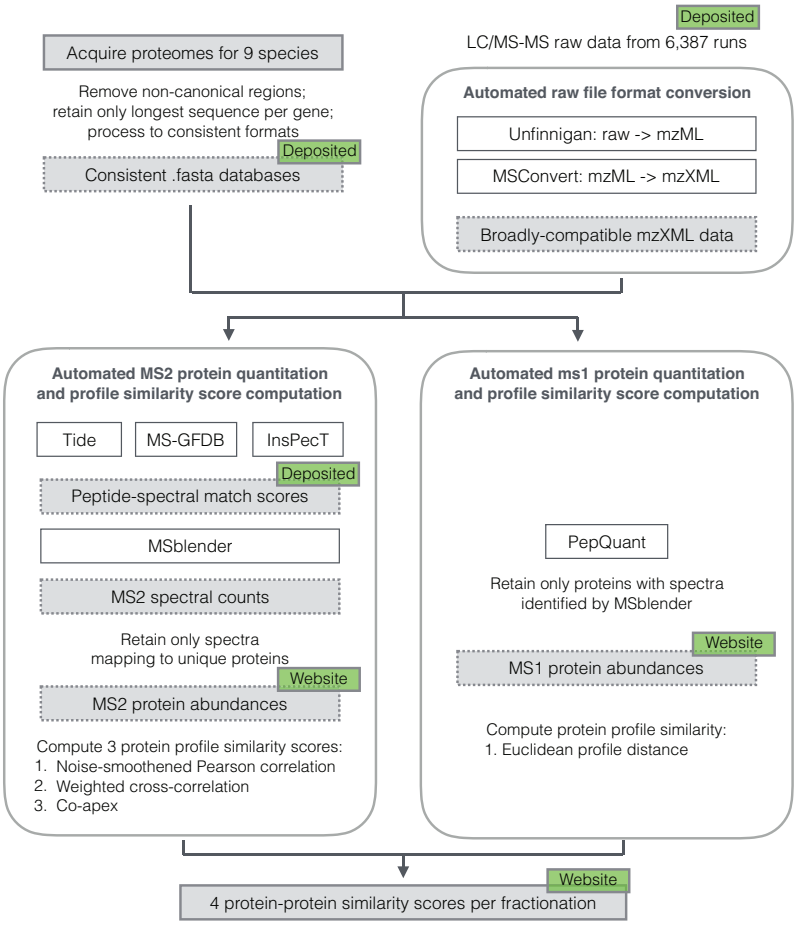


Figure 2.1: **Automation pipeline for processing data.** Top left: Proteome databases were acquired and processed to produce consistent databases across all nine species. Top right: LC/MS-MS data was acquired mostly by collaborators, organized and converted as described in the methods into a format compatible with open source proteomics tools. Databases and mzXML files were processed to produce ms1 and ms2 protein quantitations separately, and four similarity scores were computed for each fractionation from protein quantitations. Intermediate data is shaded and marked with dotted outlines; existing tools are unshaded boxes. Data marked as "Deposited" is available via PRIDE/ProteomeXchange, and data marked as "Website" is available on the project website. For further details on all these items, see Section 3.5.

species, sequence databases were processed in several ways. While the assignment of an identified peptide to multiple proteins has little detrimental effect for the purposes of most proteomics surveys, we found that such multiple assignments can be problematic to determining interactions and inferring complexes. Thus, I retained only the longest protein sequence associated with each ENSEMBL gene ID to reduce potential spurious predicted interactions between protein isoforms given their limited unique sequence information and possible biochemical similarity and to also support cleaner orthology mappings across many species. In addition, sequences associated with mutant and condition-specific chromosomal locations were removed for a similar rationale. Finally, I employed the standard method of constructing combined target-decoy databases through reversal of target protein sequences to quantify false discovery rates for peptide-spectral matches (Elias and Gygi, 2007).

2.2.4 Integrating multiple database searches

As discussed above, I exploited the added coverage and recall provided by probabilistically combining results from multiple sequence database search engines using MSblender (Kwon et al., 2011). We searched our mzXML mass spectrometry data files against our known protein databases using three established search engines using different identification and confidence assignment methodologies, Tide (Diament and Noble, 2011), MSGFDB[3], and InsPecT[4], selecting peptide-spectral matches with a false-discovery rate $<1\%$ for each sample. Further processing of

[3] `http://proteomics.ucsd.edu/Software/MSGFDB/`
[4] `http://proteomics.ucsd.edu/Software/Inspect/`

peptide and protein identifications reduced spurious associations between proteins with high sequence similarity. For example, in the case of close homologs, only peptides mapping to unique proteins were retained. While this significantly reduced the overall number of identified peptide-spectral matches, the remaining matches, still numbering more than 10 million peptide spectral matches across all experiments, were sufficient to provide high coverage across the proteomes of the various species. This additional filtering of matches dramatically reduced the occurrence of inferred associations between close homologs, which were likely spurious. See Section 2.5 for a further discussion.

2.2.5 Automation infrastructure

In order to execute the above process using the massive computational resources available at TACC, I developed a set of bash scripts, python scripts, and configuration files, with the full source code available for automated MS2 quantification on the web[5]. Together, these scripts integrate database preparation, file organization, database searching using three search engines, integration probabilistic peptide identification with MSblender, protein identification, and protein profile similarity computation (described in Chapter 3) as illustrated (Figure 2.1). Scripts for integrating with TACC's interface for parallel computation are also included to enable execution using hundreds or thousands of cores in parallel.

[5] `https://github.com/marcottelab/blendomatic`

2.2.6 MS1 for additional quantitation information

While MS2 quantitation is fast and widely used for proteome quantitation, there is some debate as to whether MS1 provides more reliable quantitation, although it may be the case that MS1 at the very least provides additional useful information (Krey et al., 2014). Thus, for the five species used to derive the conserved co-complex interactions, I further quantified MS1 precursor ion intensities as a means to improve quantification accuracy using PepQuant (Wan, Liu, et al., 2013). Initial results from PepQuant MS1 protein quantitation suggested high levels of false positive protein identifications compared to both MSblender MS2-based and MaxQuant MS1- and MS2-based ("iBAQ") quantitations (Figure 2.2), but MS1-based quantitation levels from PepQuant could provide useful additional data. Thus, we reduced false positive identifications in PepQuant by retaining only those proteins previously identified in a given sample using the MS2 spectra results derived from MSblender.

2.3 Results

In order to execute the above process using the massive computational resources available at TACC, I developed a set of bash scripts, python scripts, and configuration files, with the full source code available for automated MS2 quantification on the web[6]. Together, these scripts integrate database preparation, file organization, database searching using three search engines, probabilistic peptide

[6] `https://github.com/marcottelab/blendomatic`

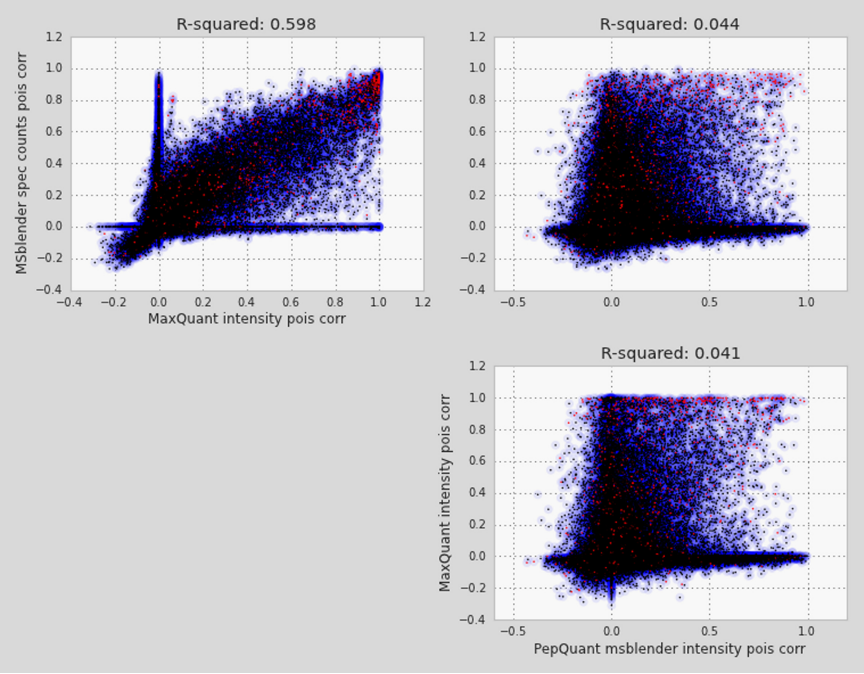


Figure 2.2: **Consistency between MSblender, PepQuant and MaxQuant.** MS1-based protein profile correlations acquired using PepQuant were compared against correlations derived using MSblender with MS2 only (our automated approach) and MaxQuant with MS1 and MS2 (a leading Windows-based software package for quantitation). While reasonably high consistency was observed between MSblender and MaxQuant results (top-left), lower consistency was observed between PepQuant and both MSblender and MaxQuant (top-right and bottom-right, respectively). Each point in a given scatter plot corresponds to a single protein's similarity score derived from the indicated quantitation method. For consistency, noise-smoothened Pearson correlation scores were used in all cases.

identification with MSblender, protein identification, and protein profile similarity computation – described further in Chapter 3—as illustrated (Figure 2.1). Scripts for integrating with TACC's interface for parallel computation are also included to enable execution using hundreds or thousands of cores in parallel.

I successfully applied this automated pipeline to our multi-species mass spectrometry data, extracting more than 10,000,000 peptide spectral matches across all experiments, corresponding to quantitations for more than 13,000 human proteins and their orthologs from 6,387 mass spectrometry experiments spanning 69 fractionations across nine species (Figure 2.3).

2.4 Discussion

Given the power of proteomics experiments to the extract the state of biological samples at the protein level, yielding an output much closer than mRNA expression patterns and with increased physiological relevance, it is not surprising that such experiments have increased dramatically in both frequency and size in the past decade. What is surprising to many is the lack of open source and scalable software tools to support such research. While the publically available pipeline described here is far from fulling addressing this need, it takes a significant step in that direction and points the way forward for future proteomics experts and software developers to build and share tools that ease the computational burden on proteomic research labs.

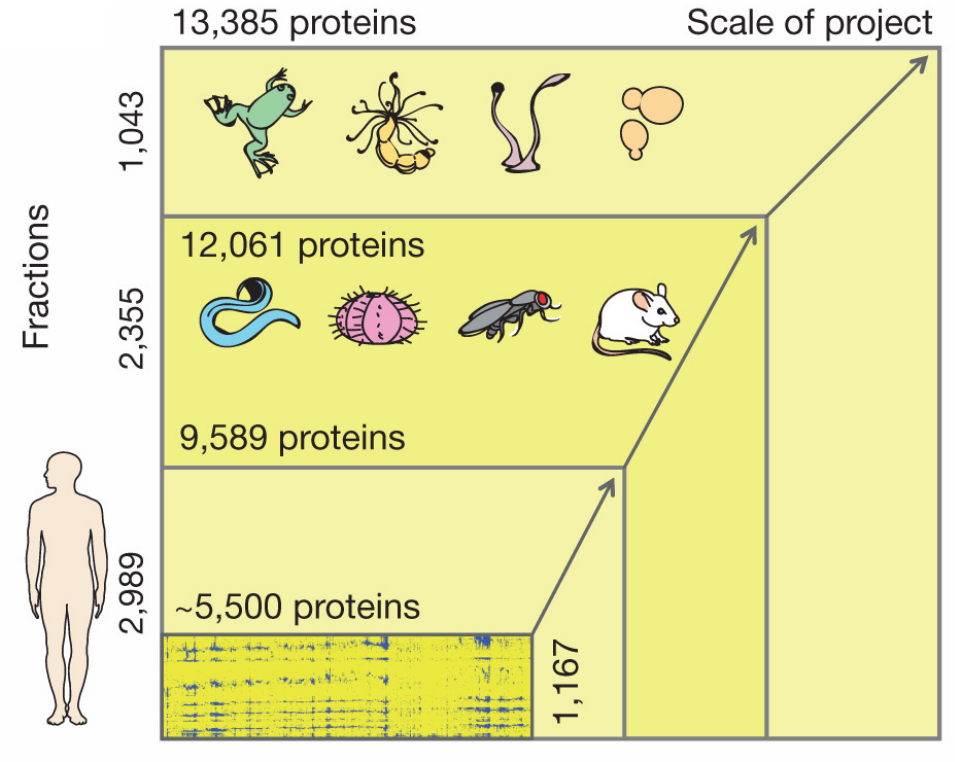


Figure 2.3: **Scale of data.** Expanded coverage via experimental scale-up relative to our prior human-only co-fractionation study (Havugimana et al., 2012). Chart shows number of proteins detected, most (63%) in two or more species. Adapted from (Wan, Borgeson, et al., 2015).

2.5 Methods

Protein databases were obtained from ENSEMBL when available (*H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, all downloaded 2012-02-10, release 65; *S. cerevisiae*, downloaded 2012-09-03, release 68) and otherwise from species-specific databases (*S. purpuratus*: spbase.org 2012-09-03, *D. discoideum*: dictybase.org 2012-02-10, *N. vectensis*: genome.jgi-psf.org 2013-02-19). Because gene models for *X. laevis* were not yet finalized, we employed interim gene models for the analysis using the released transcriptome-derived gene models (Mayball version) provided by the International *Xenopus laevis* genome project at the project website[7].

2.6 Open science

As the acquisition and processing of such a large set of proteomics experiments is a highly valuable and much needed resource to the scientific community, we made the data publicly available. Our extremely large set of supporting raw biochemical fractionation mass spectrometry data, including protein sequence databases and peptide-spectral match scores, totalling over 2TB, were deposited at ProteomeXchange. Nine datasets, each corresponding to a single species, are available using identifiers PXD002319 through PXD002328, sequentially.

As mentioned above, all scripts and full source code I developed are publicly available online[8].

[7] `http://www.marcottelab.org/index.php/Xenopus_Genome_Project`
[8] `https://github.com/marcottelab/blendomatic`

Chapter 3

Machine learning to identify conserved co-fractionation interactions

This chapter describes my use of machine learning to identify co-complex interactions using protein abundance data from fractionation experiments. Cuihong Wan and Sadhna Phanse together performed analysis describing the proportion of interactions across species, expression pattern similarity of interacting proteins, and projection of interactions across many species. Kevin Drew led analysis of the spatial relationship between interacting proteins, and collaborated with me to illustrate the ability of fractionation data to reveal complex substructure. The machine learning pipeline, the collection and integration of external data, and all other computational work involved in identifying the core conserved interactions described here were my own work, along with analyses describing the accuracy of identified interactions, contributions of features, and validation of interactions using held-out test species. Portions of the text and figures of this chapter are adapted from a paper published in Nature, where I was joint first author with Cuihong Wan[1]. I coordi-

[1]Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, Xiong X, Kagan O, Kwan J, Bezginov A, Chessman K, Pal S, Cromar G, Papoulas O, Ni Z, Boutz DR, Stoilova S, Havugimana PC, Guo X, Malty RH, Sarov M, Greenblatt J, Babu M, Derry WB, Tillier ER, Wallingford JB, Parkinson J, Marcotte EM, Emili A, Panorama of ancient metazoan macromolecular complexes. Nature, 525(7569):339-44(2015). See text above for my contributions.

nated the computational efforts, and Cuihong coordinated experimental efforts in the Emili lab.

3.1 Background

The development of co-fractionation brought a huge increase in efficiency to identifying thousands of co-complex protein-protein interactions in human cell lines (Havugimana et al., 2012). This advance stimulated a massive proteomics effort to dramatically improve the quality and coverage of co-complex interactions conserved across evolution. Previous cross-species interactome comparisons show limited overlap (Gandhi et al., 2006; Von Mering et al., 2002) due to relying on experimental data from different sources and methods. So, we sought to produce a more comprehensive and accurate map of common protein complexes using a standardized approach for multiple species. In addition to humans, we selected eight additional species for study based on their relevance as model organisms spanning roughly a billion years of evolutionary divergence (Figure 3.1). The Emili and Marcotte labs together performed more than 50 experiments involving fractionation followed by mass spectrometry, using multiple tissues and cell types in each of the nine species and employing several fractionation methodologies and a number of variants of the affinity bead type (Wan, Borgeson, et al., 2015, Supplementary Table 1). The resulting co-fractionation data acquired for *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fly), *Mus musculus* (mouse), *Strongylocentrotus purpuratus* (sea urchin), and human were used to discover conserved interactions, while the data obtained for *Xenopus laevis* (frog), *Nematostella vectensis*

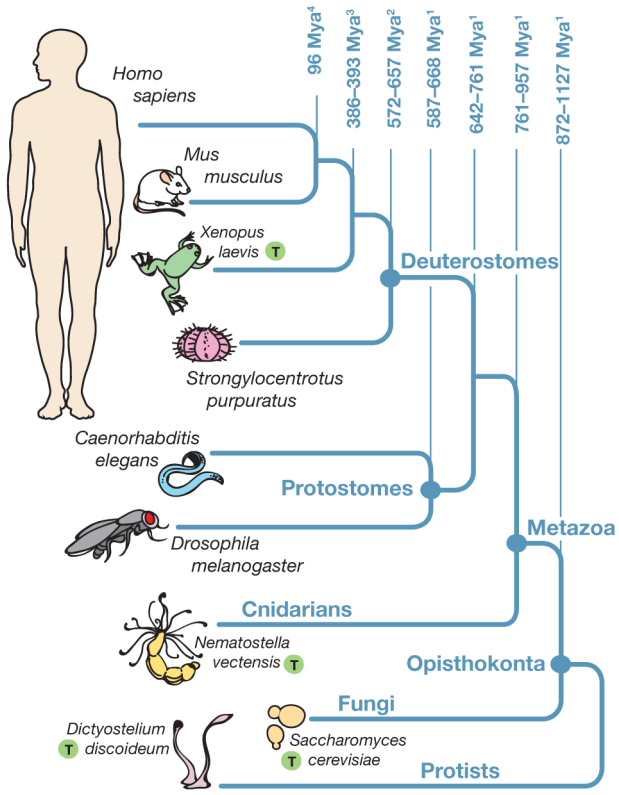


Figure 3.1: **Cross-species deep proteome fractionation.** Phylogenetic relationships of organisms analyzed in this study. We fractionated soluble protein complexes from worm (*C. elegans*) larvae, fly (*D. melanogaster*) S2 cells, mouse (*M. musculus*) embryonic stem cells, sea urchin (*S. purpuratus*) eggs and human (HEK293/HeLa) cell lines. Holdout species ('T', for test) likewise analysed were frog (*X. laevis*), an amphibian; sea anemone (*N. vectensis*), a cnidarian with primitive eumetazoan tissue organization; slime mold (*D. discoideum*), an amoeba; and yeast (*S. cerevisiae*), a unicellular eukaryote. Adapted from (Wan, Borgeson, et al., 2015).

(sea anemone), *Dictyostelium discoideum* (amoeba) and *Saccharomyces cerevisiae* (yeast) were used for independent validation.

Co-fractionation exploits the insight that components of protein complexes should elute together under varying biochemical fractionation conditions, such that proteins localized within the same complex should have higher than expected overlap in their elution profiles. This is illustrated in cartoon form and with the well-studied exocyst complex as an example in Figure 3.2.

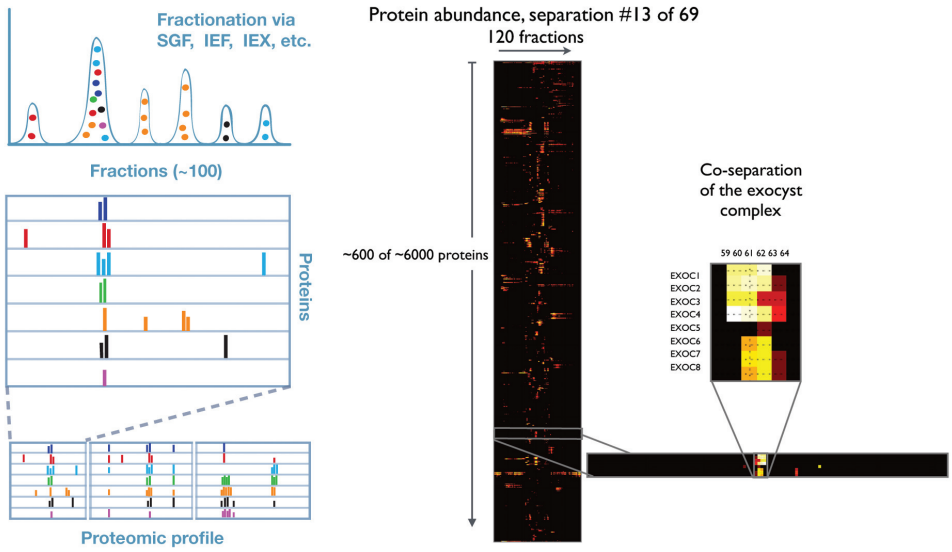


Figure 3.2: **Protein fractionation profiles.** Proteins occurring within the same fraction in a fractionation result in correlated fractionation profiles, at left. Applying multiple types of fractionations to multiple tissues and cell types extracts richer information on the consistency or context-dependence of particular interactions and sub-components. Shown at right are protein profiles in heatmap form for a portion of a representative single fractionation experiment, illustrating the co-elution of the well-studied exocyst complex in a concentrated set of three fractions near the middle of the 120 fractions in the experiment shown. Adapted from (Wan, Borgeson, et al., 2015).

The co-elution of this and many other known complexes grouped individually in this single fractionation illustrates the principle of the technique and also validates our decision to employ multiple variants of fractionation across multiple cell types and tissues. Despite the convincing degree of co-elution in such an experiment, a single experiment can result in elution profiles that overlap to such a large degree that extraction of individual complexes becomes very difficult except for a small number of cases.

3.2 Approach

In order to discover protein co-complex interactions prospectively from protein fractionation profiles across multiple species, I adopted a machine learning framework similar in concept to that employed in our lab's previous work using only human data. In this framework, every pair of observed proteins is first scored for similarity between the proteins' fractionation profiles in every fractionation experiment and across the entire dataset. Next, protein pairs exceeding a protein profile similarity threshold are scored for possible co-complex interactions using supervised machine learning with a gold standard derived from manually curated protein complexes.

3.2.1 Exploration of orthology mapping strategies

To determine the largest set of co-complex interactions from data spanning multiple species that would be of the highest value to the research community, I evaluated several approaches. The first option was to identify interactions for

each species, followed by a comparison of these independently-constructed interaction maps. This approach has obvious value to provide species-specific maps to researchers studying various model organisms, above and beyond the goal of comparing these interaction maps to discover principles of evolution and proteome organization. However, this approach also brings significant challenges, especially with the paucity of data for certain species in our captured fractionation experiments and the scant biological knowledge embedded in previously published datasets and public databases for certain species. Such species-species disparities, among other difficulties, makes false negatives a concern for any cross-species comparisons. Distinguishing the true lack of interaction from failure to observe an existing interaction in a species with a relatively smaller amount of existing data becomes unreliable and presents a large obstacle to determining clear takeaways from cross-species comparisons. While I did not pursue this approach as my primary focus, I later generated species-specific interaction maps for several species, discussed more in Chapter 5.

Rather than constructing separate species-specific interaction maps, I incorporated data across all our chosen species together in the interaction identification process, specifically focusing our analysis to identify highly-confident interactions conserved across evolution. The first inclination was a principled approach first to identify protein orthogroups for all species under investigation and then to identify conserved interactions between orthogroups. However, this approach was limiting. In early testing, it became clear that this approach led to a significant collapse in the number of interacting entities,which lost valuable interaction information. For

example in humans, we covered a evolutionary distance of more than half a billion years between the many paralogs grouped together to construct orthogroups (see Figure 3.1). In addition, identifying interactions at the level of orthogroups introduced additional rigidity into this process. Periodically, I incorporated new species into our large and growing dataset. With interactions built around orthogroups, this new data required me to rebuild our orthogroups, which would both incur significant computational overhead and create a barrier to replicability as new species were introduced. Data from a new species would create changes to our previously identified interactions and also fundamentally alter the nodes of the network. While interactions between orthogroups are appealing, in principle, for studies of evolutionary changes and divergence, this approach precludes interpretation of the nodes in our network by biologists regardless of chosen species or model organism for study. Ultimately, this approach limits its utility and thus decreases the scientific impact of this work.

For the reasons above, I mapped scored fractionation profile similarities for each input species back to a single reference species for integration and scoring, selecting human as the reference. This choice meant that my orthology work only scaled linearly with the number of species by mapping each species to human. It also meant that as I added species, I could compare changes to our inferred interactions more directly. Finally, this approach had the benefit of presenting our interactions in the format most widely interpretable, as interactions between human proteins.

3.2.2 Mapping protein profile similarity scores to human

Mapping similarities to human required selecting a method for calculating pairwise orthologies between the other eight species and human. Orthology mapping continues to attract research and discussion and is far from a solved problem, with multiple different algorithms providing different inferred orthologs. Research groups continue to evaluate the most commonly-used methods to better understand their robustness against missing data (Dalquen et al., 2013) and their degree of overlap among other topics of investigation (Erik L.L. Sonnhammer et al., 2014). I calculated pairwise orthologies between species using InParanoid 4.1 and using default settings for BLASTP with BLOSUM62 (Remm, Storm, and Erik LL Sonnhammer, 2001). In order to estimate the extent to which alternative orthology assignments may influence complex discovery, I measured the extent to which OMA-identified single copy orthologs between humans and worm, fly, and yeast, were likewise found by InParanoid. 94+% of OMA identifications were also identified by InParanoid, with only modest discordance (3–6% depending on the species) between the algorithms. This value provides an approximate bound on the extent to which alternate ortholog calls might affect co-complex discovery.

To integrate the biochemical fractionation data obtained from the other species for a given human protein pair, I selected the maximum interaction score for each species between any orthologs detected for the first human protein and any orthologs of the second human protein. In order to minimize spurious associations resulting from the "fanning out" of interaction evidence across paralogs, I required an observation of at least one correlation score greater than 0.25 in our human

fractionation experiments in order to allow the inclusion of interaction data from additional orthologs.

3.2.3 Generation of similarity scores and features

Based on the MS2-based protein spectral counts, I calculated three measures of protein co-fractionation. For a pair of protein fractionation profiles, I measured the: (1) Pearson correlation coefficient, with added Poisson noise to reduce the influence of low-count proteins ("noise-smoothened Pearson correlation"), (2) weighted cross-correlation, and (3) co-apex score, which calculated Pearson correlation and co-apex scores with Python as described previously (Havugimana et al., 2012) and employed an existing R implementation for weighted cross-correlation. I derived a fourth measure of co-fractionation from the MS1-based precursor ion peptide intensity measurements, which was implemented as 1 minus the Euclidean distance between a pair of elution profiles and calculated using pdist (SciPy Python library Oliphant, 2007). For a given protein pair, a separate score was calculated for each biochemical fractionation experiment, providing four features for each of the 55 fractionations from the input species, or 220 features total, which served as the biochemical fractionation feature inputs to the machine learning classifier described below.

3.2.4 Incorporation of external genomic and proteomic evidence

A driving factor behind the power of fractionation to discover interactions is its ability to identify interactions between any pair of proteins from a single

set of experiments rather than requiring targeted experiments for each protein, as is the case for AP-MS. A challenge that occurs with searching a huge space for possible interactions like this is the increased potential for false positives. To address this concern, I included data from other biological datasets likely to predict co-complex interactions into the machine learning pipeline as additional features. As in our previous work (Havugimana et al., 2012), an additional 19 lines of evidence were incorporated from HumanNet (Lee et al., 2011), a network of functional association between human genes derived from high- and low-throughput experimental databases from multiple species. Datasets include protein-protein association scores assembled from co-expression, domain co-occurrence, gene neighborhoods, co-inheritance, previously published high-throughput affinity purification mass spectrometry (AP-MS) and yeast two-hybrid experimental results reported prior to 2014 for human, worm, fly and yeast. Computationally predicted interactions from co-citation and literature-curated interactions were used in fly, worm, and in certain cases yeast, but excluded in human to reduce circularity, since the curated list of human gold standard complexes was derived using only low-throughput techniques described below. In addition to data from HumanNet, interactions observed in two additional recent high-throughput AP-MS studies in fly (Guruharsha et al., 2011) and human (Malovannaya et al., 2011) were also incorporated as additional features for machine learning (see Section 3.5).

3.2.5 Curation of gold-standard complexes

For supervised learning of protein co-complex interactions, I generated a gold standard reference set of complexes from a high-confidence set of manually curated mammalian protein complexes exclusively from low-throughput experiments, maintained in CORUM (Ruepp et al., 2008). I retained only annotated human complexes and eliminated approximately 10% of complexes based on them having been identified using lower confidence identification methods (see Supplementary Table 2 of Wan, Borgeson, et al., 2015 for a complete list). Complexes with more than 50 annotated members were removed to avoid bias. The resulting set of complexes was split into two sets. The first was used to train protein co-complex membership predictions and select parameters for clustering (see Chapter 4). The second set was withheld as an independent test set for the final evaluation of overall platform performance. Within each subset, positive interactions were generated from protein pairs sharing complex memberships, and negatives were generated from protein pairs contained in that positive set that did not share complex memberships in any CORUM complexes.

3.2.6 Machine learning to predict conserved interactions

The experimentally derived scores and external scores described above were used as input features to a machine learning classifier. Protein interactions were filtered before classification to include only protein pairs with direct biochemical experimental support (exceeding a score threshold of 0.5) observed in at least two or more animal species.

Performance of protein co-complex membership prediction using the more than 240 input features was evaluated using cross-validation employing only the training/cross-validation split from the gold standard reference complexes. The classifier was then trained using the entire combined training/cross-validation and test splits, and each protein pair with correlated separation profiles in fractionations from at least two species was scored by a machine learning classifier using the top 100 selected features. I evaluated multiple classifiers, and selected an SVM classifier with a linear kernel from the SVC library of Scikit-learn (Pedregosa et al., 2011), as it performed comparably to or better than other kernels and other classifiers, including ensemble classifiers, such as random forests, to predict co-complex memberships and generally produce protein interaction scores that led to better performance at the clustering stage, described in Chapter 4.

3.2.7 Evaluation of feature contributions

In order to better understand the contributions of species and data sources contributing to our identified interactions, I also utilized the inherent feature importance representation in a random forest classifier to produce the relative scores depicted in Table 3.1. This table shows the top 30 of the more than 240 total features that served as input to the pairwise machine learning interaction scoring pipeline. This ranking illustrates the breadth of species, external data sources, and fractionation types that proved most valuable in terms of predictive power against our gold standard manually curated human co-complex interactions. Since several of the most important features were from external datasets, I examined how sensitive our

recovery of gold standard interactions was to the lack of the features ranked as most important from the external fly AP-MS dataset (Guruharsha et al., 2011). The removal of even this most important feature reduced recall at most by a few percent at any level of precision.

Table 3.1: Features with highest calculated importance in predicting interactions

Score	Species	E for External	Details
0.218	Fly	E	AP-MS
0.186	Human	E	AP-MS
0.099	Human		Hela nuclear; Isoelectric focusing; MS2 co-apex scoring
0.058	Yeast	E	Co-expression
0.058	Yeast	E	AP-MS
0.049	Yeast	E	Literature-curated
0.047	Human		Hela nuclear; Heparin; MS2 weighted cross-correlation scoring
0.047	Human		Hela cytosolic; Triple-phase isoelectric focusing; MS2 weighted cross-correlation scoring
0.038	Worm		Heparin; MS1 euclidean distance scoring
0.038	Human		Hela cytosolic; Sucrose gradient; MS2 pearson correlation scoring
0.038	Worm		Whole worms; Heparin; MS2 weighted cross-correlation scoring
0.032	Human		Hela cytosolic; Isoelectric focusing; MS2 pearson correlation scoring
0.031	Sea urchin		Isoelectric focusing; MS2 weighted cross-correlation scoring
0.029	Human		Hela nuclear; Sucrose gradient; MS2 weighted cross-correlation scoring
0.029	Fly	E	Physical interactions
0.026	Human		Hela cytosolic; Triple-phase isoelectric focusing; MS2 weighted cross-correlation scoring
0.026	Human		Neural stem cells; Heparin; MS2 pearson correlation scoring
0.025	Mouse		Embryonic stem cells; Heparin; MS1 euclidean distance scoring
0.024	Yeast	E	Co-citation
0.023	Human		HEK nuclear; Heparin; MS2 pearson correlation scoring
0.023	Worm		Whole worms; Affinity separation flow-through; MS2 weighted cross-correlation scoring
0.023	Human		HEK; Agilent HPLC; MS2 weighted cross-correlation scoring
0.022	Human		Hela nuclear; Triple-phase isoelectric focusing; MS2 weighted cross-correlation scoring
0.021	Worm		Heparin; MS1 euclidean distance scoring
0.021	Worm	E	Co-citation
0.021	Human		HEK nuclear; Heparin; MS2 weighted cross-correlation scoring
0.021	Worm		Whole worms; Affinity separation; MS1 euclidean distance scoring
0.021	Worm		Whole worms; Affinity separation; MS2 pearson correlation scoring
0.021	Human	E	Phylogenetic co-inheritance
0.02	Worm		Whole worms; Affinity separation; MS2 weighted cross-correlation scoring

3.3 Results

3.3.1 A high-confidence network of co-complex interactions

The machine learning pipeline described above identified and quantified 13,386 protein orthologs across 6,387 fractions obtained from 69 different experiments, an order of magnitude expansion in data coverage relative to the prior human-only previous study (Havugimana et al., 2012). Individual pair-wise protein associations were scored based on the fractionation profile similarity measured in each species. Measurements of overall performance showed high precision with reasonable recall by the co-fractionation data alone (Figure 3.3), with external data sets serving only to increase precision and recall, as all derived interactions were required to have biochemical support. Co-fractionation data of each input species affected overall performance, in each case increasing precision and recall (Figure 3.4). The final filtered interaction network consisted of 16,655 high-confidence co-complex interactions in human, available for download[2]. All interactions were supported by direct biochemical evidence in at least two input species, with half (8,121) detected in three or more (Figure 3.5). This enabled cross-species modelling and functional inference.

3.3.2 Functional and disease enrichment of interacting proteins

Multiple lines of evidence support the quality of the network. Reference complexes withheld during training were reconstructed with higher precision and recall (Figure 3.3) relative to the previous human-only map (Havugimana et al.,

[2]"PPI & Correlations" at `http://metazoa.med.utoronto.ca`

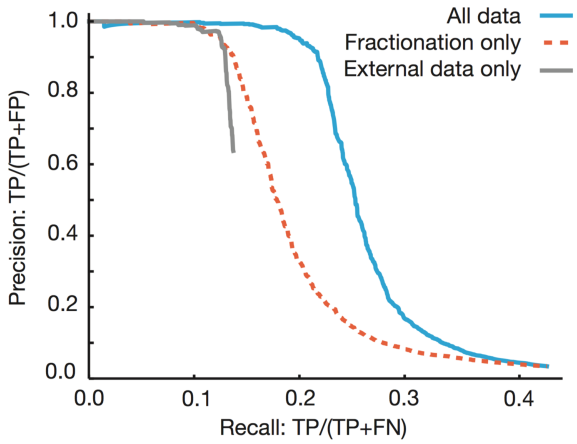


Figure 3.3: **Relative contributions of fractionation and external data.** Performance benchmarks, measuring precision and recall of our method and data in identifying known co-complex interactions (annotated human complexes from CORUM [Ruepp et al. 2008]). Complexes were split into training and withheld test sets; five-fold cross-validation against 4,528 interactions derived from the withheld test set shows strong performance gains, beyond baselines achieved using only co-fractionation or external evidence alone. TP, true positive; FP, false positive; FN, false negative. Adapted from (Wan, Borgeson, et al., 2015).

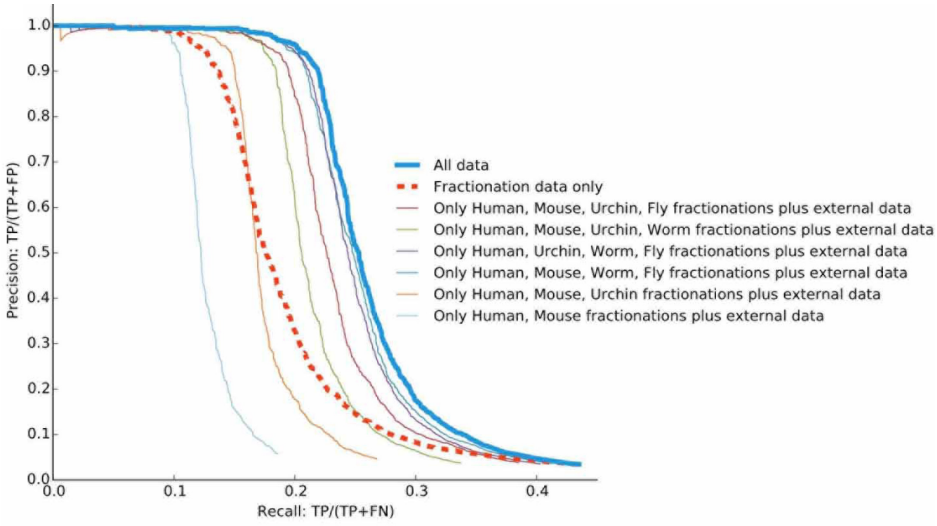


Figure 3.4: **Contributions of data from additional species.** Performance benchmarks, measuring the precision and recall of our method and data in identifying known co-complex interactions from a withheld reference set of annotated human complexes (from CORUM [Ruepp et al. 2008]; as in Figure 3.3). Five-fold cross-validation against this withheld set shows strong performance gains, beyond a baseline achieved using only human and mouse co-fractionation data along with additional evidence from independent protein interaction screens (Guruharsha et al., 2011; Malovannaya et al., 2011) and a functional gene network (Lee et al., 2011) (far-left curve), made by integrating co-fractionation data from the additional non-human animal species (as indicated). "All data" and "Fractionation data only" curves include biochemical fractionation data from all five input species: human, mouse, urchin, fly and worm; the latter curve omits all external data. In all cases, at least two species were required to show supporting biochemical evidence. Recall refers to the fraction of 4,528 total positive interactions derived from the withheld human CORUM complexes. Adapted from (Wan, Borgeson, et al., 2015).

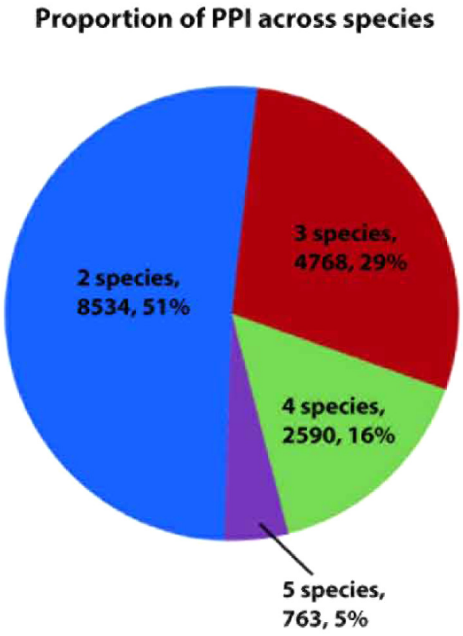


Figure 3.5: **Proportion of PPI across species.** All 16,655 interactions were identified at least in two species, half (49%, 8,121) found in three or more species. Adapted from (Wan, Borgeson, et al., 2015).

2012). The interacting proteins were also enriched sixfold (hypergeometric $P < 1 \times 10^{-24}$) for shared subcellular localization annotations in the Human Protein Atlas Database (Uhlén, Oksvold, et al., 2010), 21-fold enriched ($P < 1 \times 10^{-56}$) for shared disease associations in OMIM22 and were highly correlated with human tissue proteome abundance profiles (Kim et al., 2014) (Figure 3.6).

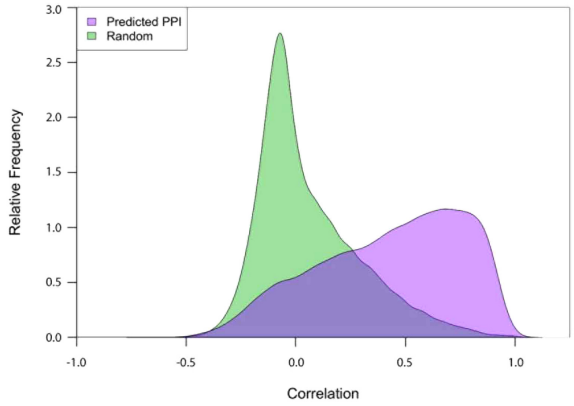


Figure 3.6: **Expression pattern similarity of interacting proteins.** Distribution of global protein tissue expression pattern similarity, measured as the Pearson correlation coefficient of protein abundance across 30 human tissues (Kim et al., 2014), showing markedly higher correlations for 16,468 pairs of putative co-complex interaction partners compared to the same number of randomized pairs of proteins in the network which were not predicted to interact. Adapted from (Wan, Borgeson, et al., 2015).

3.3.3 Inferred structural relationships through hierarchical protein profile similarities

Besides indicating stably associated proteins, our multispecies biochemical profiles faithfully recapitulated the architecture of multiprotein complexes of known three-dimensional structure, with a general trend for most correlated protein pairs

to be spatially closer (Figure 3.7). For example, hierarchical clustering of 30S

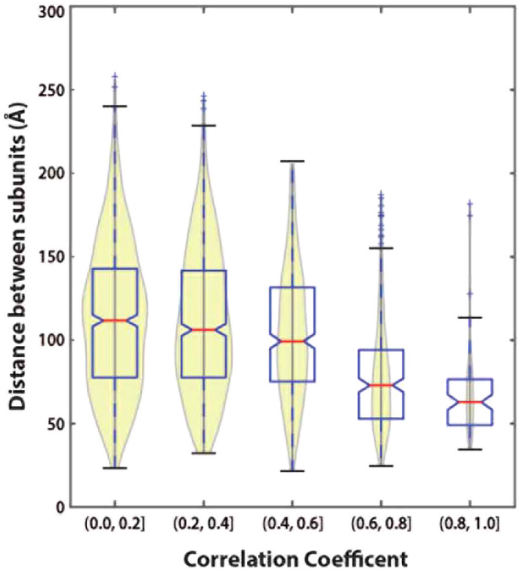


Figure 3.7: **Highly correlated proteins are spatially closer.** The degree of co-fractionation is measured as the correlation coefficient between elution profiles. Spatial proximity is calculated from the mean of residue pair distances between components of multisubunit complexes with known three-dimensional structures. Adapted from (Wan, Borgeson, et al., 2015).

proteasome subunits according to chromatographic elution profiles of all five input species correctly separated the 20S and 19S particles and the regulatory lid from the base sub-complex (Figure 3.8), reflecting known hierarchies of complex formation and disassembly.

3.3.4 Verification of the conserved nature of identified interactions

To independently verify the reliability of these projections, I examined the co-fractionation profiles of putatively interacting orthologs (interologs) in the four

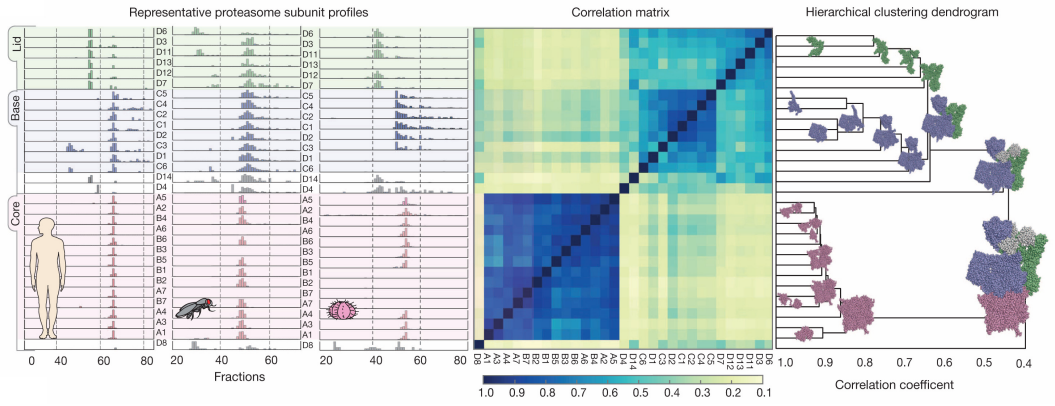


Figure 3.8: **Hierarchical clustering reveals proteasome substructure.** Left, representative co-fractionation data (normalized spectral counts shown for portions of 3 of 42 experimental profiles) from human, fly and sea urchin showing characteristic profiles of proteasome core, base and lid sub-complexes. Hierarchical clustering (right) of pan-species pairwise Pearson correlation scores (centre) is consistent with accepted structural models (Protein Data Bank ID: 4CR2; core, red; base, blue; lid, green; out-clusters, white). Adapted from (Wan, Borgeson, et al., 2015).

holdout species, as obtained by protein quantification across 1,127 biochemical fractions (see Section 3.5). Most of the predicted interactors showed highly correlated co-fractionation profiles among the holdout test species to a degree comparable to those of the input species used for learning (Figure 3.9). The biochemical

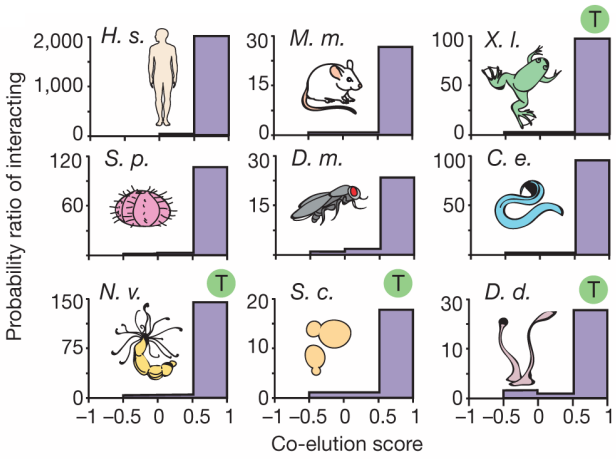


Figure 3.9: **Conservation of interactions in independent species.** Plots showing high enrichment (probability ratio of interacting) of predicted interacting orthologous protein pairs (relative to non-interacting pairs) among highly correlated fractionation profiles, in both the holdout validation (test, T) and input species (colours reflect clade memberships). Adapted from (Wan, Borgeson, et al., 2015).

data obtained for sea urchin and sea anemone showed slightly better agreement than that for Dictyostelium and yeast, which was proportional to evolutionary distance (Rubin et al., 2000).

3.3.5 Prediction of conserved interactions across species

Because most interacting components were phylogenetically conserved across vast evolutionary timescales, our conserved interactions were further used to predict

over one million high-confidence co-complex interactions among orthologous protein pairs for 122 extant eukaryotes with sequenced genomes, available for download[3]. The number of interactions ranged from ~8,000 to ~15,000 per species depending on the phyla (Figure 3.10), with more projected among deuterostomes,

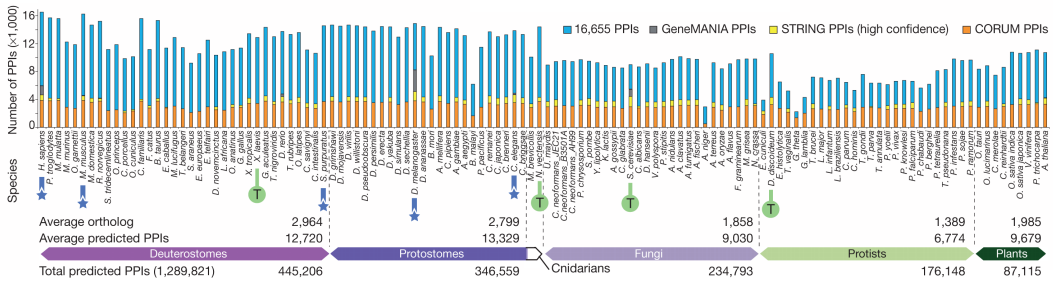


Figure 3.10: **Projection of conserved co-complex interactions across 122 species.** Projection of conserved co-complex interactions across 122 eukaryotic species, indicating overlap with leading public PPI reference databases. STRING bars indicate excess over CORUM; GeneMANIA bars indicate excess over both; component and interaction occurrences across clades indicated at bottom. Adapted from (Wan, Borgeson, et al., 2015).

protostomes and cnidaria, and showing high component retention. Fewer were projected in fungi, plants and, especially, protists, where the relative paucity of co-complex conservation probably reflects inherent clade diversity, especially in parasite genomes (for example, gene loss among Apicomplexa). While largely congruent with previous smaller-scale studies of PPI conservation (Bezginov et al., 2013), the majority of conserved co-complex interactions are novel, as less than one-third are curated in CORUM, STRING and GeneMANIA databases. This markedly

[3]"Predicted PPI" at `http://metazoa.med.utoronto.ca`

increased the number of metazoan protein interactions reported to date, covering roughly 10%–25% of the estimated conserved animal cell interactome (Stumpf et al., 2008; G. Traver Hart, Ramani, and Edward M. Marcotte, 2006) and promotes many new avenues of inquiry.

3.4 Discussion

Accurate identification of physiologically relevant protein interactions has long been a challenging undertaking. Here, using a large number of co-fractionation experiments using different fractionation methods on different cell types and tissues from many animal species and incorporating additional evidence of interactions, I identified and validated more than 13,000 conserved interactions to a level of confidence that allowed their projection across over a hundred species, taking a large step forward in the number of confident interactions expected across an important branch of the tree of life. In addition, among member proteins of the same complex, I identified a hierarchical structure within the fractionation profiles signatures and determined that they indicated complex subcomponents, proximity and assembly order. These results together illustrate the value of integrating multiple methods of biochemical fractionation to support the identification of protein interactions and interrogation of complex substructure.

3.5 Methods

3.5.1 Filtering away of peptides matching to multiple proteins

As discussed in Chapter 2, both our protein sequence databases and our peptide-spectral matches were processed to reduce the occurrence of spurious assignment of a single peptide to multiple proteins. In fact, from the onset I used only the single longest protein sequence per gene. The later decision also removed peptide-spectral matches that could correspond to multiple distinct genes, an artifact apparent in our investigations of inferred co-complex interactions. These interactions initially contained numerous examples of inferred interactions between genes that were clearly close homologs–in fact a significant fraction of observed co-complex interactions fell into this category, which was an unexpected observation. By removing all peptide-spectral matches corresponding to peptides with perfect matches against multiple proteins, I observed the near complete removal of this observed effect. While it is possible that some or even many of these inferred interactions truly occur in the underlying biology, inferring interactions with high sensitivity between highly similar proteins using mass spectrometry is a challenge for additional experimental methods or more nuanced computational approaches to examine.

3.5.2 Incorporation of external interaction datasets

While the fly data (Guruharsha et al., 2011) was conveniently available in a format easily interpretable to use in inferring co-complex interactions, human data (Malovannaya et al., 2011) required additional processing. For each pairwise in-

teraction indicated by the human dataset, the authors indicated confidence by naming each inferred complex ("MemoID") with the first letter A/P/T indicating the confidence level of accepted, provisional, and tentative, respectively. In surveying the distribution of these confidence levels and some cross-validation evaluation, I scored each interaction as 10 for each accepted interaction, 3 for provisional and 1 for tentative.

3.5.3 Validation using co-fractionation of independent taxa

To validate that the 16,655 high-confidence co-complex interactions were in fact conserved across species and not used to generate the map, I determined whether and to what degree interacting pairs of proteins in our high confidence network were more likely than non-interacting pairs of proteins to co-fractionate together in the four independent test species—the African clawed frog *X. laevis*, the starlet sea anemone *N. vectensis*, the budding yeast *S. cerevisiae*, and the multicellular amoeba *D. discoideum*. For the set of non-interacting proteins, I took the 3,464 proteins forming our high-confidence interaction network and formed all possible pairs between them not accounted for in our interaction data, amounting to 5,981,261 putatively non-interacting pairs. For each of the four independent species, I calculated the Pearson correlation coefficient with added Poisson noise to reduce the influence of low-count proteinsthe first of the four co-fractionation scores described above, and the one with the broadest coverage across pairs. To integrate biochemical fractionation data for each of the four independent species for a given human protein pair, the maximum interaction score was taken between any

orthologs of the first human protein and any orthologs of the second human protein. These scores were averaged across all fractionations for each species respectively. For each species, I then tallied the number of interacting and non-interacting pairs scoring between -1 and -0.5, .-5 and 0, 0 and .5, and .5 and 1, and then calculated a ratio with the fraction of interacting pairs scoring in that bin as the numerator and the fraction of non-interacting pairs scoring in that bin as the denominator, analogous to the calculation of relative risk in epidemiological studies.

3.6 Open science

High-confidence conserved interactions in human are available for download[4]. Python code for the entire machine learning pipeline, including training/test set creation from the gold standard, cross-validation, and supporting functions to generate many of the visualizations, is publicly available online[5].

[4] "PPI & Correlations" at `http://metazoa.med.utoronto.ca`
[5] `https://github.com/marcottelab/infer_complexes`

Chapter 4

Clustering and exploration of protein complexes

This final component of my work describes the development and use of clustering metrics and methods to derive protein complexes from co-fractionation protein interactions. Experimental validation and functional exploration of the Commander complex were performed by Fan Tu in John Wallingford's developmental biology lab. Cuihong Wan and Sadhna Phanse in the Emili lab analyzed evolutionary age and abundance trends among complexes. Kevin Drew led analysis of molecular weights. The exploration of clustering methods and metrics and development of the automated clustering pipeline were my own work, along with generation of the identified conserved protein complexes and global computational validation. I analyzed sequential metabolic enzyme complexes as suggestions of possible metabolic channeling. Portions of the text and figures of this chapter are adapted from a paper[1] published in Nature, on which I was joint first author, coordinating the computational efforts, with Cuihong Wan, who coordinated the biological experimental efforts in the Emili lab.

[1]Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, Xiong X, Kagan O, Kwan J, Bezginov A, Chessman K, Pal S, Cromar G, Papoulas O, Ni Z, Boutz DR, Stoilova S, Havugimana PC, Guo X, Malty RH, Sarov M, Greenblatt J, Babu M, Derry WB, Tillier ER, Wallingford JB, Parkinson J, Marcotte EM, Emili A, Panorama of ancient metazoan macromolecular complexes. Nature, 525(7569):339-44(2015). See text above for my contributions.

4.1 Background

Pairwise protein complex co-membership scores from the machine learning pipeline described above formed a weighted protein interaction network. A useful way to interpret such networks is through a soft partition into stable assemblies of multiple proteins into complexes (Alberts, 1998). Low-throughput studies and manually curated protein interactions often take this approach (Ruepp et al., 2008). Therefore, I developed a method to cluster our network of highly confident protein interactions into such a representation that identified dense clusters to reveal protein complexes.

4.1.1 Clustering graphs

Clustering graphs has been a topic of some research for several decades. With the advent of high-throughput protein-protein interaction studies, clustering protein interaction networks has seen increasing research efforts from many groups over the last 15 years. The Markov Clustering Algorithm (MCL) is particularly prominent among general purpose graph clustering algorithms still in wide use to cluster protein interactions. MCL is based on the principles of flow and random walks, with the simple idea that a random walk visiting a densely connected region in a graph will not leave that densely connected region as often, indicating a cluster (van Dongen, 2000). Its continued use is likely due, at least in part, to its broad applicability, ease of use, easy-to-visualize algorithm and computational efficiency.

4.1.2 Clustering protein interactions

In the following decade, numerous approaches were developed specifically with the goal of clustering protein interactions, including MCODE (Bader and Hogue, 2003) and CFinder (Adamcsek et al., 2006). MCODE introduced a different algorithm along with a number of additional parameters into the clustering process to permit researchers to hone size, shape, and other aspects of identified complexes. It also introduced appropriate metrics to evaluate the quality of protein clusters. Rather than partitioning a graph into subgraphs according to density, CFinder introduced the idea of overlapping complexes, in which nodes (proteins) could participate in multiple clusters. This approach necessitates developing different algorithms, but closely aligns with the understanding most biologists hold regarding complex identification. More recently, a group of collaborators from our previous human fractionation project (Havugimana et al., 2012) evaluated a host of protein interaction clustering tools and developed a method, called ClusterONE, to address their perceived shortcomings in the existing approaches (Nepusz, Yu, and Paccanaro, 2012). The group developed three new clustering metrics, which they combined, to represent their best representative estimate of clustering quality. They found MCL generally produced the best scores across several different protein interaction datasets. They sought to meet or exceed its evaluated performance while maintaining the ability found in some new algorithms to allow proteins to participate in multiple complexes. They succeeded in this goal for their chosen metrics, with ClusterONE producing consistently higher clustering quality than MCL, while retaining many features introduced by MCODE, and also supporting the assignment

of proteins to multiple complexes.

4.2 Approach

4.2.1 An automated pipeline for clustering

In our initial attempts to identify protein complexes from our co-fractionation network of protein interactions, I employed ClusterONE given its advantages discussed in the introduction. As mentioned, ClusterONE employs several parameters useful for tuning the properties of identified complexes to drive better performance depending on the structure and size of the input interaction network. To explore the range of possible parameter choices, I built an automated pipeline that tested the output of the clustering process against a training set of complexes derived from CORUM, our gold standard of manually-curated human complexes (Ruepp et al., 2008). I also tested MCL, verifying that ClusterONE produced complexes scoring better against the gold standard while allowing protein assignment to multiple complexes.

4.2.2 Opportunities for clustering improvement

However, while visually inspecting high-scoring clusterings, I found that the metrics employed previously by the ClusterOne authors did not sufficiently account for many features of clustering that we found important to the quality of the output. One particular disregarded aspect of the clustering output was a measure of redundancy among the identified complexes. If proteins can participate in multiple complexes, the opportunity exists for certain combinations of clustering parameters

to result in clusterings that achieve high scores per the ClusterONE developers' chosen metrics, while exhibiting a high degree of redundancy, for instance, representing both a 10-unit complex along with several highly similar 8- or 9-unit versions.

A second disadvantage of these initial high-scoring clusterings was a perceived over-collapse of multiple very large complexes. Identifying a true complex significantly larger than the largest ones known, such as the ribosome, would be remarkable, but highly surprising. So, I was skeptical when I discovered multiple such candidate complexes in the early highest-scoring overall clusterings, especially as investigation into the numerous members left little room for doubt that multiple complexes, connected by numerous lower-confidence interactions, were being grouped together. In essence, in a parameter regime performing well on the high number of small- to medium-sized complexes, ClusterONE performed poorly on very large complexes in our view. While ClusterONE's metrics concerning the quality of the overall clustering were strong, the trade-off in poor quality on a small number of larger complexes was unsatisfactory. So, I sought to develop a combination of new metrics and different methods to resolve the issue.

4.2.3 Exploration of clustering quality and metrics

Given the dissatisfaction with the redundancy and over-collapsed nature observed in the tested ClusterONE results, I explored numerous additional clustering metrics suggested in prior work (Bader and Hogue, 2003; Adamcsek et al., 2006), in addition to developing several new metrics. These metrics were developed primarily to evaluate a proposed clustering of protein interactions against the training split

of gold standard complexes derived from CORUM. I also devised several metrics simply to measure aspects of the clustering itself, such as the level of redundancy. Initial attempts to improve overall clustering quality involved using various combinations of selected metrics to choose some proposed clusterings. Each had the highest combined score along some particular combination of metrics when compared to the training set of gold standard complexes. Their performance on each of the 15 metrics is illustrated in Figure 4.1.

Optimizing against such a large number of clustering metrics and understanding the various trade-offs to balance metrics against one another proved challenging, so progress toward an enhanced clustering was elusive. In my attempts to manage these trade-offs, I noticed that a number of the metrics fluctuated together, which implied that the correlation structure among many metrics could provide a useful simplification of the task at hand. Surprisingly, a simple hierarchical clustering of the metrics based on pairwise correlations of their values across a few thousand clusterings from my automated parameter exploration revealed that the clustering metrics fell into essentially two groups—one group correlating with higher recall and another group correlating with higher precision (Figure 4.2). The strong correlation structure enabled me to pare down to just three crucial metrics.

From this point onward, I evaluated clustering performance using only three measures: the average non-overlap, maximum matching ratio, and sensitivity. The average non-overlap was calculated by averaging the number of clusters in the map over a range of thresholds from 0 to 1, for which another cluster could be identified yielding a pairwise overlap greater than the given threshold. This average

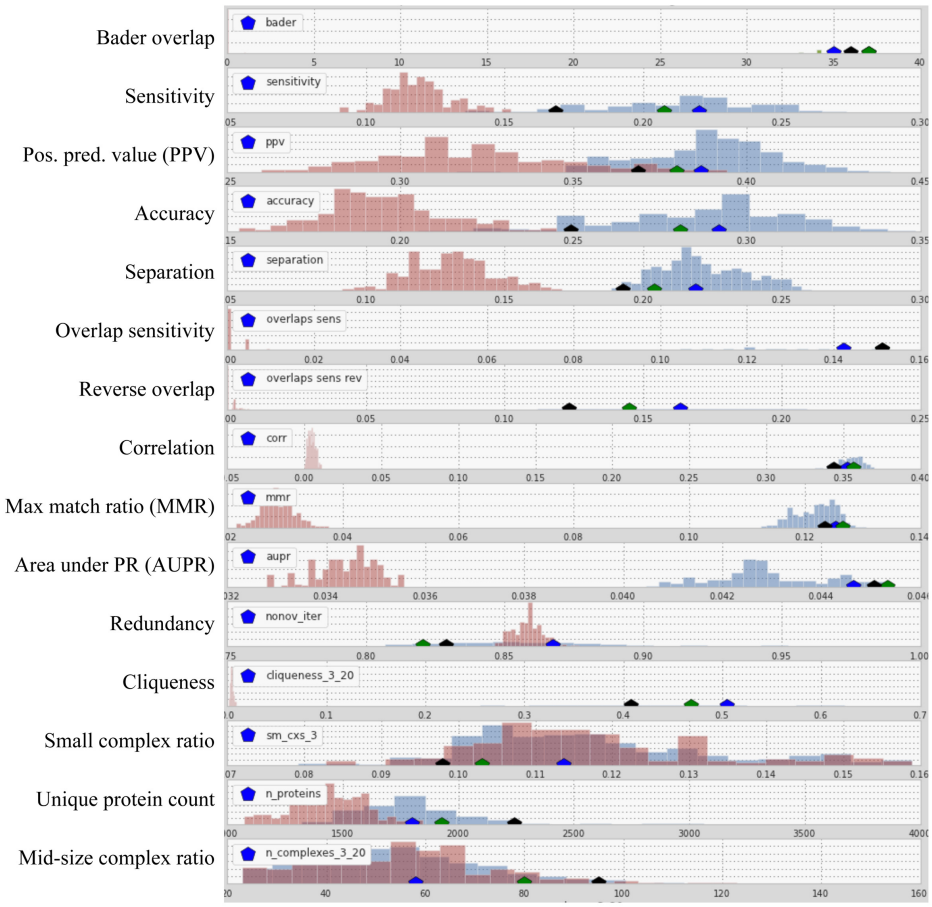
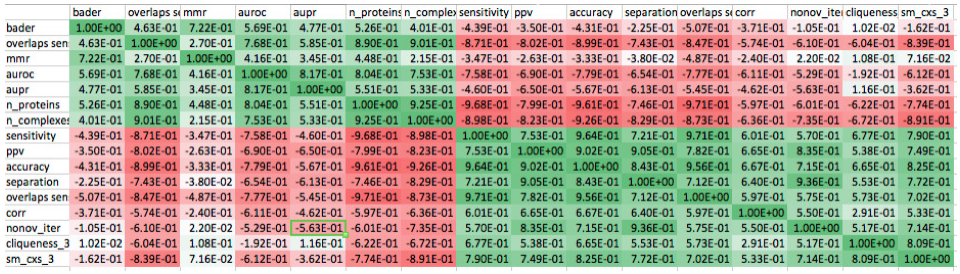


Figure 4.1: **Optimizing clustering against many metrics.** These 15 histograms correspond to 15 different clustering metrics, named at left. Data comes from over 1,000 clusterings, 900 produced by the automated clustering pipeline (shown in blue) and 200 produced by random permutations of randomly-chosen clusterings to preserve overall network topology (shown in red). The blue, green, and black markers represent 3 different candidate clusterings. The blue clustering is the tightest, with the highest precision as measured against the gold standard corum complexes, and has the highest scores on metrics related to precision, like separation (fifth from the top), and lowest on scores that reflect the overall coverage of the map, like n_proteins (total number of unique proteins, second from the bottom). The green is medium, and the black is lowest precision and highest coverage.



	bader	overlaps s	mmr	auroc	aupr	n_proteins	n_complex	sensitivity	ppv	accuracy	separation	overlaps s	corr	nonov_ite	cliqueness	sm_cxs_3
bader	1.00E+00	4.63E-01	7.22E-01	5.69E-01	4.77E-01	5.26E-01	4.01E-01	-4.39E-01	-3.50E-01	-4.31E-01	-2.25E-01	-5.07E-01	-3.71E-01	-1.05E-01	1.02E-02	-1.62E-01
overlaps sen	4.63E-01	1.00E+00	2.70E-01	7.68E-01	5.85E-01	8.90E-01	9.01E-01	-8.71E-01	-8.02E-01	-8.99E-01	-7.43E-01	-8.47E-01	-5.74E-01	-6.10E-01	-6.04E-01	-8.39E-01
mmr	7.22E-01	2.70E-01	1.00E+00	4.16E-01	3.45E-01	4.48E-01	2.15E-01	-3.47E-01	-2.63E-01	-3.33E-01	-3.80E-02	-4.87E-01	-2.40E-01	2.20E-02	1.08E-01	7.16E-02
auroc	5.69E-01	7.68E-01	4.16E-01	1.00E+00	8.17E-01	8.04E-01	7.53E-01	-7.58E-01	-6.90E-01	-7.79E-01	-6.54E-01	-7.77E-01	-6.11E-01	-5.29E-01	-1.92E-01	-6.12E-01
aupr	4.77E-01	5.85E-01	3.45E-01	8.17E-01	1.00E+00	5.51E-01	5.33E-01	-4.60E-01	-6.50E-01	-5.67E-01	-6.13E-01	-5.45E-01	-4.62E-01	-5.63E-01	1.16E-01	-3.62E-01
n_proteins	5.26E-01	8.90E-01	4.48E-01	8.04E-01	5.51E-01	1.00E+00	9.25E-01	-9.68E-01	-7.99E-01	-9.61E-01	-7.46E-01	-9.71E-01	-5.97E-01	-6.01E-01	-6.22E-01	-7.74E-01
n_complexes	4.01E-01	9.01E-01	2.15E-01	7.53E-01	5.33E-01	9.25E-01	1.00E+00	-8.98E-01	-8.23E-01	-9.26E-01	-8.29E-01	-8.73E-01	-6.36E-01	-7.35E-01	-6.72E-01	-8.91E-01
sensitivity	-4.39E-01	-8.71E-01	-3.47E-01	-7.58E-01	-4.60E-01	-9.68E-01	-8.98E-01	1.00E+00	7.53E-01	9.64E-01	7.21E-01	9.71E-01	6.01E-01	5.70E-01	6.77E-01	7.90E-01
ppv	-3.50E-01	-8.02E-01	-2.63E-01	-6.90E-01	-6.50E-01	-7.99E-01	-8.23E-01	7.53E-01	1.00E+00	9.02E-01	9.05E-01	7.82E-01	6.65E-01	8.35E-01	5.38E-01	7.49E-01
accuracy	-4.31E-01	-8.99E-01	-3.33E-01	-7.79E-01	-5.67E-01	-9.61E-01	-9.26E-01	9.64E-01	9.02E-01	1.00E+00	8.43E-01	9.56E-01	6.67E-01	7.15E-01	6.65E-01	8.25E-01
separation	-2.25E-01	-7.43E-01	-3.80E-02	-6.54E-01	-6.13E-01	-7.46E-01	-8.29E-01	7.21E-01	9.05E-01	8.43E-01	1.00E+00	7.12E-01	6.40E-01	9.36E-01	5.53E-01	7.72E-01
overlaps sen	-5.07E-01	-8.47E-01	-4.87E-01	-7.77E-01	-5.45E-01	-9.71E-01	-8.73E-01	9.71E-01	7.82E-01	9.56E-01	7.12E-01	1.00E+00	5.97E-01	5.75E-01	5.73E-01	7.02E-01
corr	-3.71E-01	-5.74E-01	-2.40E-01	-6.11E-01	-4.62E-01	-5.97E-01	-6.36E-01	6.01E-01	6.65E-01	6.67E-01	6.40E-01	5.97E-01	1.00E+00	5.50E-01	2.91E-01	5.33E-01
nonov_iter	-1.05E-01	-6.10E-01	2.20E-02	-5.29E-01	-5.63E-01	-6.01E-01	-7.35E-01	5.70E-01	8.35E-01	7.15E-01	9.36E-01	5.75E-01	5.50E-01	1.00E+00	5.17E-01	7.14E-01
cliqueness_3	1.02E-02	-6.04E-01	1.08E-01	-1.92E-01	1.16E-01	-6.22E-01	-6.72E-01	6.77E-01	5.38E-01	6.65E-01	5.53E-01	5.73E-01	2.91E-01	5.17E-01	1.00E+00	8.09E-01
sm_cxs_3	-1.62E-01	-8.39E-01	7.16E-02	-6.12E-01	-3.62E-01	-7.74E-01	-8.91E-01	7.90E-01	7.49E-01	8.25E-01	7.72E-01	7.02E-01	5.33E-01	7.14E-01	8.09E-01	1.00E+00

Figure 4.2: **Redundancy among clustering metrics.** Shown is the correlation between each pair of metrics across thousands of clusterings, green highlighting a positive correlation, and red negative. The correlations show that the metrics break down very strongly into two camps: the top/left ones that reflect or are tied mainly to the size and recall of the clusterings, and the bottom/right ones that approximately represent the precision/specificity of the clusterings.

value was then subtracted from one to provide a positive, increasing measure of quality. Next, the maximum matching ratio assigns one-to-one best matches between complexes in the gold standard and predicted sets. It reports the average overlap score achieved for gold standard complexes using the same overlap score as above. Finally, sensitivity scores the precision of the predicted complexes against the gold standard, both as previously described (Havugimana et al., 2012; Brohée and Helden, 2006). Empirically, I found that requiring a minimum value of 0.75 for average non-overlap for a clustering produced maps with a reasonably low level of redundancy according to myself and colleagues, while still allowing complexes to overlap enough to be biologically feasible. Within this constraint, clusterings were selected to maximize performance according to the maximum matching ratio and sensitivity metrics.

4.2.4 Two-stage clustering

While the winnowing of metrics to three and the added constraint involving the new non-overlap metric simplified clustering evaluation and resolved issues with redundancy, the over-collapse of large complexes remained a concern. As described above, visual inspection of some of the largest complexes suggested that they were composed of multiple potentially high-quality smaller complexes, bound together with numerous, but somewhat weak, interactions between them. Attempts to repair this issue using different parameter choices in ClusterONE met with very limited success before having a significant negative impact on the overall quality of the rest of the identified complexes in the clustering. Recalling the effectiveness of MCL in effectively cutting graphs into smaller subgraphs, I experimented with a two-stage clustering process. Complexes identified via ClusterONE were immediately processed individually with MCL. By merging MCL and its single tunable parameter into my clustering parameter exploration pipeline, I produced clusterings that scored higher on all metrics, while simultaneously resolving the over-collapse of the largest complexes and retaining a relatively low level of redundancy. See Section 4.5 for more details on optimal clustering parameters.

4.2.5 Incorporating non-clustered high-scoring interactions

While the two-stage clustering process described above successfully resolved concerns with over-collapse and redundancy, one new concern arose regarding the robustness of the clustering process to input interactions. I found that even very slight alterations to the scores of the determined protein interactions often

caused significant fluctuations in the complexes identified by the clustering process, in particular the appearance or disappearance of numerous single interactions. To increase the robustness of the clustering, I included all very highly confident interactions, roughly 500, in the final clustering output, described further in Section 4.5.

4.3 Results

4.3.1 A map of conserved metazoan complexes

The 981 putative multiprotein groupings include both well-known and novel complexes linked to diverse biological processes (Figure 4.3[2]). The final conserved metazoan complexes comprise 7,669 high confidence pairwise interactions among 2,153 distinct (human) proteins. The complexes have estimated component ages spanning from ~500 million (metazoan-specific or "new") to over one billion years (ancient or "old") of evolutionary divergence. Co-complex interactions among the complexes show an improvement in precision relative to the high-confidence interactions that served as input to the clustering process (Figure 4.4).

4.3.2 Independent biological assessment

We used multiple approaches to assess the accuracy of the predicted complexes. First, my colleagues performed affinity purification mass spectrometry (AP/MS) experiments on select novel complexes based on their absence from curated databases BioGrid v3.2.102 (Stark et al., 2006) and IRefWeb v4.1 (Turner

[2]Available for download under "Complex Network" at `http://metazoa.med.utoronto.ca`

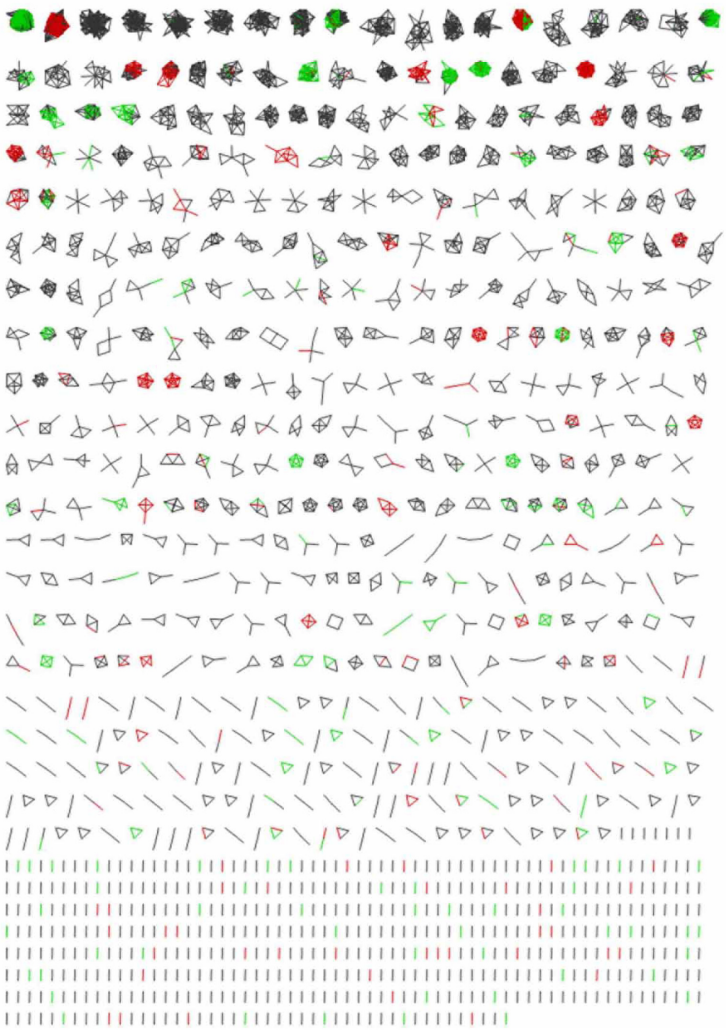


Figure 4.3: **981 conserved animal protein complexes.** Schematic of 981 identified complexes containing 2,153 unique proteins. In this graphical representation, 7,669 co-complex interactions are shown as lines, and proteins as nodes. Red and green interactions were previously annotated in CORUM. Red interactions were used in training the classifier and/or clustering procedure, while green interactions were held out for validation purposes. Grey interactions were not previously annotated in CORUM. Adapted from (Wan, Borgeson, et al., 2015).

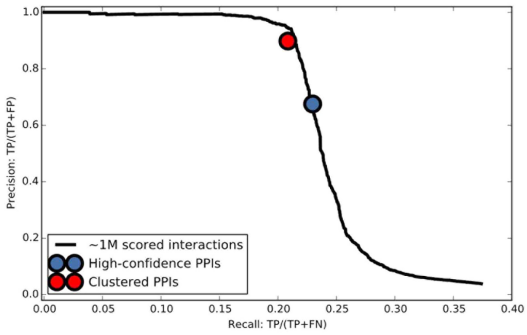


Figure 4.4: **Final precision/recall performance on withheld interaction test set.** A support vector machine classifier was trained using interactions derived from our training set of CORUM complexes, then, 1 million protein pairs found to co-elute in at least two of the five input species were scored by the classifier. Black curve shows precision and recall for ranked list of co-eluting pairs, with recall representing the fraction recovered of 4,528 total positive interactions derived from the withheld set of merged human CORUM complexes, and precision measured using co-eluting pairs where both members of the pair are contained in the set of proteins represented in the CORUM withheld set. The top 16,655 pairs, giving a cumulative precision of 67.5% and recall of 23.0% on this withheld test set, form the high- confidence set of co-complex proteinprotein interactions (blue circle). Adapted from (Wan, Borgeson, et al., 2015).

et al., 2010) which indicated that they were not reported in either human, mouse, worm, fly or yeast. These experiments validated most associations in both worm and human (see Wan, Borgeson, et al., 2015). Second, I performed a global validation by comparing our derived complexes to a recent large-scale AP/MS study of 23,756 putative human protein interactions detected in cell culture (Huttlin, Lily Ting, et al., 2015) along with CORUM curated mouse complexes and CYC2008 (Finn et al., 2014) yeast complexes. In all cases, I observed a partial, but highly statistically significant, overlap to a degree comparable to literature-derived complexes (Figure 4.5).

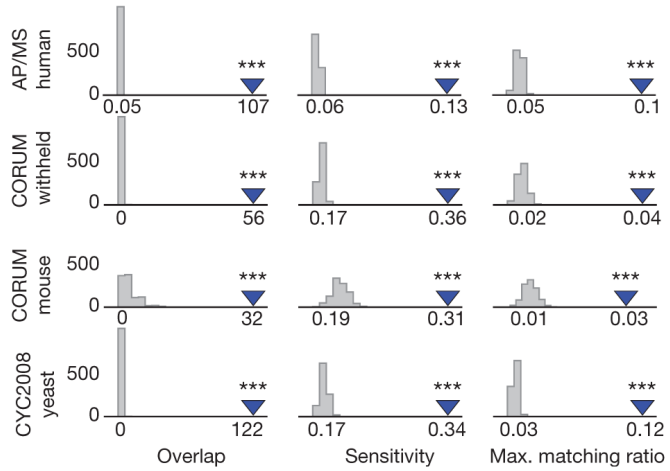


Figure 4.5: **Global validation of complexes.** Conserved complexes significantly overlap large-scale AP/MS data reported for human cell lines (Huttlin, L. Ting, et al., 2016) to a comparable extent as literature reference sets (Ruepp et al., 2008), using three measures of complex-level agreement (see Section 4.5); ***P , 0.001, determined by shuffling (grey distributions). Adapted from (Wan, Borgeson, et al., 2015).

4.3.3 Evolutionarily conserved complexes and subunits

Although proteins arising in metazoa account for about three quarters of all human gene products through gene duplication or other means, they form only 39% of the clusters (Figure 4.6, a). These "new" complexes tend to be smaller (#3 components; Figure 4.6, b) and specific to components not present in "mixed" complexes. Even though protein number and diversity greatly increased with the rise of animals (Bezginov et al., 2013), these "new" complexes suggest that most stable protein complexes were inherited from a unicellular ancestor and subsequently modified over time to various extents (Figure 4.6, c). Indeed, the dominant phylogenetic profile of complexes across Eukarya (Figure 4.6, d) is composed of ancient subunits ubiquitous among eukaryotes, either entirely (344 old complexes) or predominantly (490 mixed complexes), the latter presumably reflecting preferential accretion of additional components to pre-existing macromolecules (Eisenberg and Levanon, 2003).

These primordial complexes are present throughout the Opisthokonta supergroup (animals and fungi), estimated to be more than one billion years old (Knoll, 1992), and plants (and presumably lost/significantly diverged among parasitic protists). Reflecting this central importance, these complexes have strong, ubiquitous expression, abundant throughout all cell types and tissues (Figure 4.7).

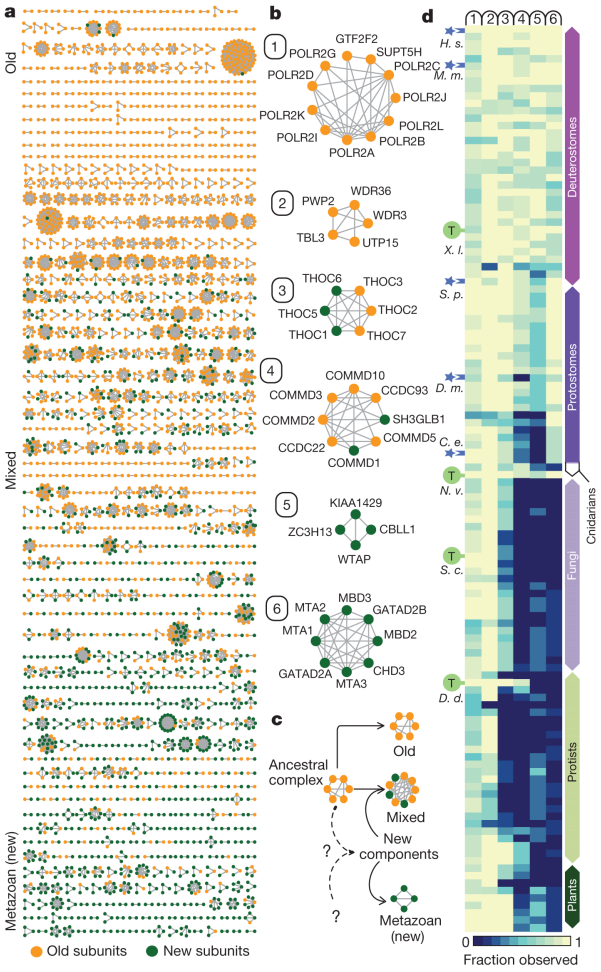


Figure 4.6: **Conservation of protein complexes across Metazoa and beyond. a,** Conserved multiprotein complexes, identified by clustering, arranged according to average estimated component age (see Bezginov et al., 2013). Proteins (nodes) classified as metazoan (green) or ancient (orange); assemblies showing divergent phylogenetic trajectories termed "mixed". **b,** Example complexes with different proportions of old and new subunits. **c,** Presumed origins of metazoan (new), mixed and old complexes; "?" indicates variable origins of new genes. **d,** Heat map showing prevalence of selected complexes across phyla. Colour reflects fraction of components with detectable orthologues (absence, dark blue). Sea anemone (*N. vectensis*) is the most distant metazoan (cnidarian) analyzed biochemically. Adapted from (Wan, Borgeson, et al., 2015).

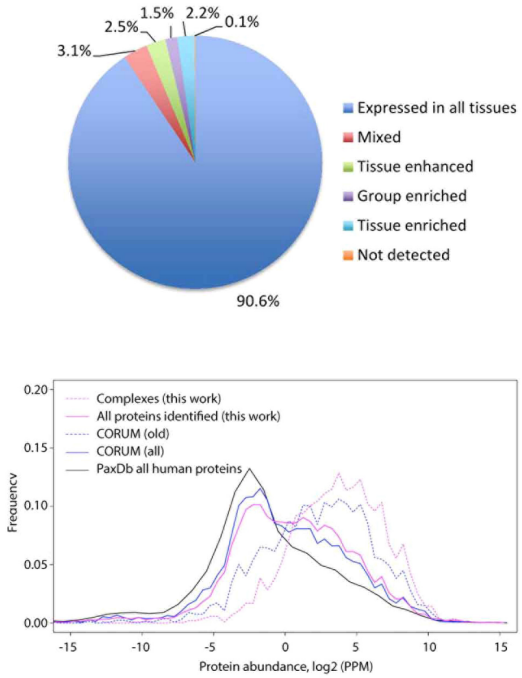


Figure 4.7: **Abundance and expression trends for proteins in complexes.** Proteins within the identified complexes tend to be ubiquitously expressed across human tissues. Top, pie chart shows the proportions of proteins with varying tissue expression patterns, from a recently published human tissue proteome map (Uhlén, Fagerberg, et al., 2015). Consistent with this observations, 91% of the protein components in the complexes were expressed in more than 15 tissues in data from a reference human proteome (Kim et al., 2014), compared to less than half (46%) of the 17,294 proteins in the overall reference set (Z-test $P < 0.001$) Bottom, The distributions of average protein (data from PaxDb integrated data set, 9606-H.sapiens_whole_organism-integrated_data set) abundances for all proteins identified and those within complexes. Evolutionarily old proteins (defined by OMA as described in Bezginov et al., 2013 and mentioned earlier) tend towards higher abundances, even for proteins in reference complexes. Adapted from (Wan, Borgeson, et al., 2015).

4.3.4 Experimental validation and functional characterization of a novel conserved complex

We also observed broad agreement between the derived complexes' inferred molecular weights (assuming 1:1 stoichiometries) and migration by size-exclusion chromatography (Figure 4.8). A prime example is the coherent profiles of a large

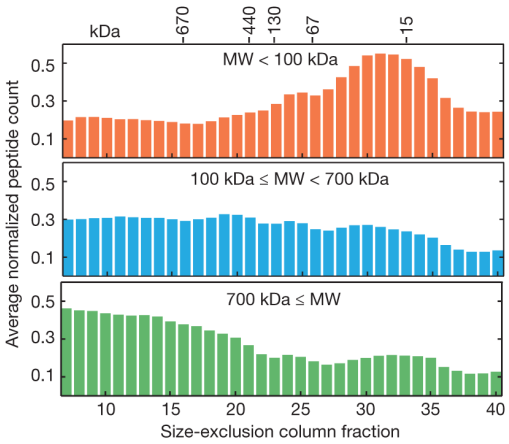


Figure 4.8: **Agreement of complex size with size-exclusion data.** Inferred molecular weights (MW) of human protein complexes tend to agree with size-exclusion chromatography profiles from Kirkwood and Lamond, 2013. Adapted from (Wan, Borgeson, et al., 2015).

(~500kDa) mixed complex with several un-annotated components (Figure 4.9), dubbed "Commander", because most subunits share COMM (copper metabolism MURR1) domains (Burstein et al., 2005) implicated in copper toxicosis (van de Sluis et al., 2002), among other roles (Burstein et al., 2005; McDonald, 2013). Commander contains coiled-coil domain proteins CCDC22 and CCDC93 in addition to ten COMM domain proteins, broadly supported by co-fractionation in

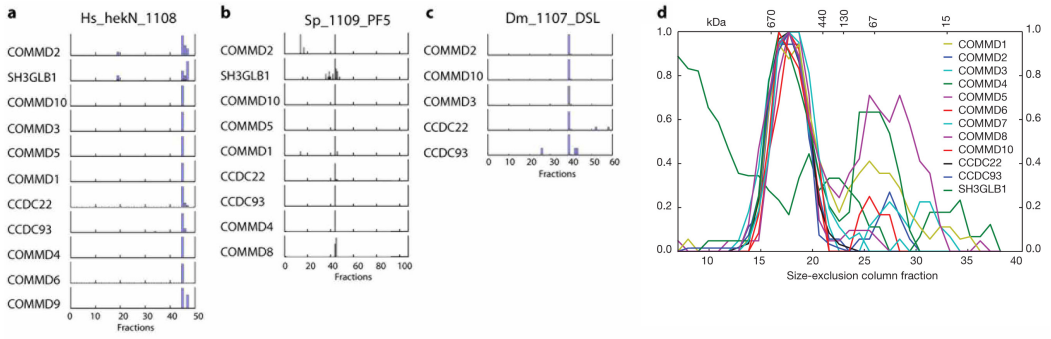


Figure 4.9: **Co-fractionation consistency of the Commander complex.** Example protein elution profiles are plotted for Commander complex subunits observed from: HEK293 cell nuclear extract (a); sea urchin embryonic (5 days post-fertilization) extract (b); and fly SL2 cell nuclear extract (c); each fractionated by heparin affinity chromatography. d, Co-elution of human Commander complex subunits by size-exclusion chromatography, consistent with an approximately 500-kDa particle. Adapted from (Wan, Borgeson, et al., 2015).

human, fly and sea urchin[3].

We found an unexpected role in embryonic development for Commander, whose subunits are strongly co-expressed in the developing frog (Figure 4.10). COMMD2/3-knockdown (morpholino) tadpoles showed impaired head and eye development (Figure 4.11) and defective neural patterning and expression changes in brain markers PAX6, EN2 and KROX20/EGR1 (Figure 4.12). Given the recently discovered link (Kolanczyk et al., 2015; Voineagu et al., 2012) between CCDC22 and human syndromes that exhibit intellectual disability, malformed cerebellum and craniofacial abnormalities, the deep conservation of the Commander complex

[3]See supporting website, `http://metazoa.med.utoronto.ca/php/view_elution_image.php?id=71&cond=ms2`

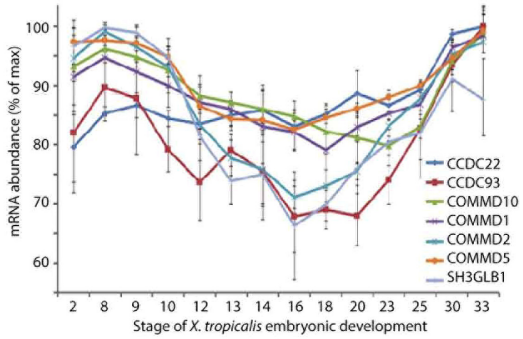


Figure 4.10: **Developmental co-expression of Commander subunits.** Co-expression of Commander complex subunits during embryonic development of *X. tropicalis* (plotting mean plus/minus standard deviation of three clutches; data from Yanai et al., 2011). Adapted from (Wan, Borgeson, et al., 2015).

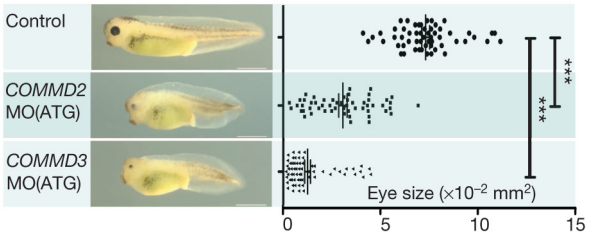


Figure 4.11: **Impaired eye development in Commander morphants.** Morpholino (MO(ATG), targeting start codon to block translation) knockdown of COMMD2 (n=55 animals, 2 clutches, 1 eye each) or COMMD3 (n=64) in *X. laevis* embryos causes defective head and eye development (control n=57). $^{***}P < 0.0001$, two-sided Mann-Whitney test. Adapted from (Wan, Borgeson, et al., 2015).

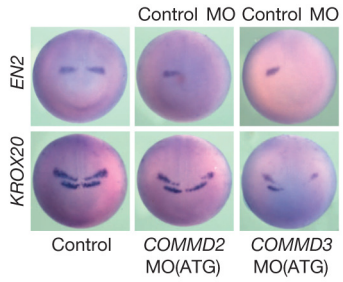


Figure 4.12: **Altered neural patterning with Commander knockdown.** COMMD2/3 knockdown animals (five embryos per treatment examined) show altered neural patterning, including posterior shift or loss of expression of mid-brain marker EN2 and KROX20 (EGR1). Adapted from (Wan, Borgeson, et al., 2015).

suggests COMMD2/3 is a strong candidate in the aetiology of these heterogeneous disorders.

4.3.5 Network perspective into conserved biological systems

Knowledge of conserved macromolecular associations provides a roadmap for additional functional inferences. For instance, fractionation profiles can be compared for any protein pair in our data set to search for evidence of interactions. I found a significant enrichment for interactions among pairs of human proteins acting sequentially in annotated pathways (Croft et al., 2014), especially G-protein and MAP-kinase cascades (Table 4.1). Enzymes acting consecutively in core metabolic reactions also showed a higher tendency to interact. The significance of these interaction tendencies decayed with more intervening steps. For example, strong consecutive interactions occurred within the widely conserved purine biosynthetic pathway with enzymes eluting in two peaks (for example, PAICS, GART), one

Table 4.1: Consecutive pathway and metabolic pairs

Gene Symbol 1	Gene Symbol 2	Cofrac_ PPI_Score	In_ Corum	Reaction Network	Reactome Reaction or Metabolic Pathway
EEF1B2	EEF1A1	1.00	0	Signaling	REACT_67.1
EEF1D	EEF1A1	1.00	0	Signaling	REACT_67.1
PDS5B	PDS5A	1.00	1	Signaling	REACT_150250.3
UBE2I	SAE1	0.85	0	Signaling	REACT_163816.3 REACT_163643.3 REACT_163963.3
EIF5	EIF2S2	0.73	0	Signaling	REACT_1060.1 REACT_656.1
EEF1A1	EEF2	0.68	1	Signaling	REACT_1937.1
EIF5	EIF3I	0.46	0	Signaling	REACT_1060.1 REACT_656.1
EIF5	EIF2S1	0.46	0	Signaling	REACT_1060.1 REACT_656.1
UBE2I	UBA2	0.37	0	Signaling	REACT_163816.3 REACT_163643.3 REACT_163963.3
RAB8A	RAB10	0.27	0	Signaling	REACT_147836.3
MAPK1	MAPK3	0.22	0	Signaling	REACT_111117.2
EIF3G	EIF5	0.20	0	Signaling	REACT_1060.1 REACT_656.1
RHOA	TRIO	0.13	0	Signaling	REACT_10098.1 REACT_19216.3
RAC1	RHOC	0.12	0	Signaling	REACT_19137.4
EIF5	EIF3B	0.11	0	Signaling	REACT_1060.1 REACT_656.1
EIF4G1	TNKS1BP1	0.09	0	Signaling	REACT_20651.2
STAM2	TSG101	0.09	0	Signaling	REACT_27272.2
PCNA	POLE	0.08	1	Signaling	REACT_353.2 REACT_677.2
HADHA	HADHB	1.00	0	Metabolic	Mitochondrial beta-oxidation of long chain fatty acids
TKT	TALDO1	0.99	0	Metabolic	Nonoxidative pentose phosphate pathway
TKTL2	TALDO1	0.97	0	Metabolic	Nonoxidative pentose phosphate pathway
PPAT	GART	0.44	0	Metabolic	Purine biosynthesis
HIBADH	ALDH6A1	0.34	0	Metabolic	Catabolism of L-valine
PHGDH	PSAT1	0.27	0	Metabolic	L-serine synthesis
UAP1	PGM3	0.27	0	Metabolic	Amino sugar and nucleotide sugar metabolism
GART	PFAS	0.27	0	Metabolic	Purine biosynthesis
HSD17B10	ACAA2	0.22	0	Metabolic	Amino acid metabolism
LDHA	ALDH2	0.16	0	Metabolic	Glycolysis and Amino acid metabolism
PYGB	GBE1	0.16	0	Metabolic	Starch and sucrose metabolism
TKT	RPE	0.14	0	Metabolic	Nonoxidative pentose phosphate pathway
LDHA	ALDH1B1	0.14	0	Metabolic	Glycolysis
PYGL	GBE1	0.13	0	Metabolic	Starch and sucrose metabolism
ACAT1	ACAA2	0.13	0	Metabolic	Fatty acid degradation,Valine, leucine and isoleucine degradation
ADSL	ATIC	0.11	0	Metabolic	Purine biosynthesis
PYGM	GBE1	0.08	0	Metabolic	Starch and sucrose metabolism

coincident with the prior enzyme and the second with the downstream enzyme, suggestive of substrate channeling (Ovádi, 1995) (Figure 4.13).

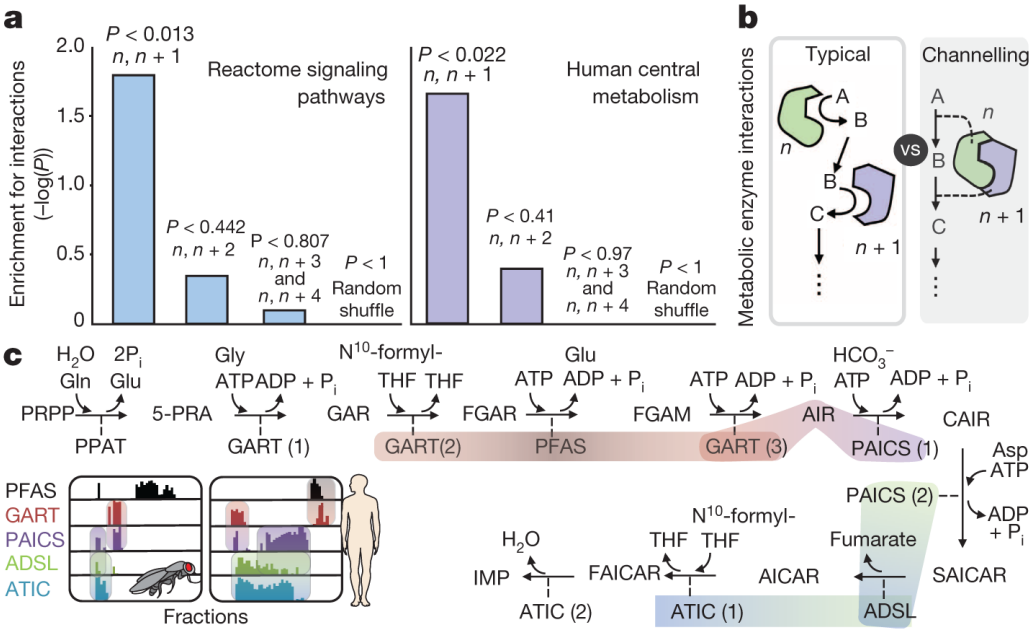


Figure 4.13: **Interactions between consecutive pathway and metabolic pairs. a**, Enrichment (permutation test P value) for interactions among sequential pathway components and metabolic enzymes relative to shuffled controls (n refers to enzyme index, where n,n+1 denotes sequential enzymes, n,n+2 sequential-but-one, and so on, as described in Section 4.5.) **b**, Metabolic channelling as opposed to traditional (typical) two-step cascade model. **c**, Conserved interactions among consecutively acting enzymes involved in purine biosynthesis (two representative co-fractionation profiles of the 69 total generated are shown). Adapted from (Wan, Borgeson, et al., 2015).

4.4 Discussion

While it might have been expected that identifying accurate and sensible protein complexes from protein interaction networks would be a replicable and straightforward process, this does not turn out to be the case. Clustering graphs is by no means a solved research problem. The various approaches previously used, when applied to this set of identified protein interactions, suffered in my hands from some combination of redundancy, high sensitivity to input interactions, lack of participation in multiple complexes, over-collapse of large complexes, and lack of accuracy against gold standard curated human complexes. Through extensive exploration with the help of the automated clustering and parameter exploration pipeline I developed, I succeeded in establishing a clustering approach that resolved the various shortcomings of prior methods. I used this new approach to derive a set of nearly 1,000 conserved animal complexes, and validated them computationally and many of them, with the help of colleagues, experimentally. These complexes have already led to many insights and discoveries, some of which are described here.

4.5 Methods

4.5.1 Clustering parameters

I found and jointly optimized the key parameters that drove clustering variation and performance. I also found a combination of parameters that satisfied the average non-overlap redundancy constraint and maximized the maximum matching ratio and sensitivity values. ClusterOne employs two main parameters—a mini-

mum density for identified complexes and a penalty term to slow agglomeration. Optimization resulted in selecting values of 0.35 and 1.0, respectively. In comparing balanced and unbalanced training sets in the scoring of protein pairs, clustering results were sensitive to the rate of falloff of these pairwise scores, so a scaling parameter was developed. and was applied in the step before clustering. Since I used an unbalanced training set for better performance, this parameter was 0.20. MCL has a single key parameter, called inflation or I, optimized to be a value of 3.0 here. Finally, the number of high pairwise protein co-complex scores provided as input to the clustering process affected the sensitivity and coverage of the resulting map of complexes. The optimal value was found to be 1.0% of the pairwise protein co-membership scores or 9,989 scored pairs. Following identification of strong and significant complexes, a larger set of high-scoring protein interactions (10% of the interactions with biochemical evidence or 99,888 pairs) were used to define the significant interactions observed within the identified complexes. These interactions form the edges in the final complex map.

4.5.2 Incorporation of high-confidence interactions

When very high confidence co-complex associations were excluded by the clustering process, I incorporated them back into the map using the following approach. First, interactions were judged high-confidence for a score $>$0.9 in the cross-validated precision-recall PPI evaluation, which yielded 283 additional interactions for inclusion in the map. Second, I re-ranked the top 20,000 interactions using a selected threshold from the precision-recall curve by the total number of

biochemical separations in which the protein pair was assigned a high (>0.5) correlation score. I selected all interactions from this set having 20 or more high-scoring fractionations, yielding 239 interactions for inclusion. The union of these two sets yielded a total of 507 additional, previously excluded interactions for inclusion in the map. A final single round of clustering using MCL was applied with the same parameters applied in stage two of the clustering process ($I = 3.0$). The resulting complexes were added to the map, including pairwise interactions excluded from larger complexes. Finally, redundant complexes were merged if either 1) the complex pair exceeded the Bader-Hogue overlap threshold of 0.55 as described above or 2) the smaller of the two complexes had more than four members in which all or all-but-one were a subset of a larger complex.

4.5.3 Analysis of consecutively acting signal transduction and metabolic enzyme interactions

In order to test comprehensively for consecutively acting, putatively interacting proteins across cellular pathways in general, proteins in signal transduction and other non-metabolic pathways were assembled from the Reactome database (Croft et al., 2014), requiring annotations associated with the term "reaction" while excluding interactors related by the terms "direct_complex" and "indirect_complex", resulting in 32,703 sequentially acting proteins in cellular pathways, including in signal transduction, e.g., G-protein and MAPK cascades. In order to ask specifically whether sequential metabolic enzymes were enriched for interactions, I defined our set of confident sequential enzymes as the intersection of sequential pairs found in processing two comprehensive human metabolic networks,

Recon2 (v02) (Thiele et al., 2013) and KEGG (Kanehisa and Goto, 2000, downloaded July 27, 2013). For Recon2, I excluded 36 common metabolites, available in Table 4.2. The intersection of these two sets of sequential interactions yielded 647 confident sequential metabolic enzymes.

Table 4.2: 36 common metabolites excluded from Recon2

ATP(3-)	glutathionate(1-)
adenosine 3',5'-bismonophosphate(4-)	H2O
ADP	hydrogen peroxide
ADP(3-)	Hydrogen peroxide
Ammonium	hydrogenphosphate
AMP	NAD(1-)
AMP(2-)	NADH(2-)
ATP(4-)	Nicotinamide adenine dinucleotide
Bicarbonate	Nicotinamide adenine dinucleotide - reduced
CMP	Nicotinamide adenine dinucleotide phosphate
CMP(2-)	Nicotinamide adenine dinucleotide phosphate - reduced
CO2	O2
Coenzyme A	proton
Diphosphate	Reduced glutathione
Flavin adenine dinucleotide oxidized	Sodium
Flavin adenine dinucleotide reduced	UDP
GDP	UDP(3-)
GDP(3-)	water

Both pathway sets (general cellular pathways and metabolic enzyme pathways) were then analyzed in the same fashion, detailed here for the metabolic case: I first asked whether these pairs of sequential enzymes were enriched for higher co-complex scores, according to the output of our integrative machine learning process of identifying interactions, compared to a reshuffled set of false interactions formed from the same set of enzymes. Using a two-sample KS test, I calculated the p-value that the distribution of interaction scores from our sequential enzyme set differed from the distribution of scores from reshuffled negative interactions, finding indeed a significant enrichment for higher scores ($p < 0.022$). I repeated the analysis for enzymes separated by two steps, by three or four steps, and by an independent

reshuffled set of enzyme pairs, finding decreasing significance of enrichment for higher scores in each case, as shown in Figure 4.13, a. Among the high-confidence interaction partners, I identified 17 examples of sequentially acting, physically interacting enzyme pairs, including six falling into two respective pathways of three pairs each: the purine biosynthetic pathway, and the pentose phosphate pathway.

The consecutively acting, co-complex Reactome pathway and Recon/KEGG enzyme pairs are listed in Table 4.1.

4.6 Open science

Conserved complexes are available for download[4]. Code for the full clustering pipeline, clustering evaluation, and generation of many figures is publicly available online[5].

[4] "Complex Network" at `http://metazoa.med.utoronto.ca`
[5] `https://github.com/marcottelab/infer_complexes`

Chapter 5

Conclusions and future directions

The sequencing of entire genomes provided the parts lists for biological research. These sequences can be compared across near and distant species to reveal evolutionary principles and to distribute knowledge gained in one species across evolutionary distances, which among other things significantly enables model organism research to directly shed light on human biology and disease. This integrated and comprehensive understanding of the function of a growing number of genes and proteins also advances our ability to interpret diseases and other differences between human individuals in terms of their genetic basis, stimulating the emerging and growing era of personalized medicine. However, this view of biology lacks mechanistic insight into the underlying factors, such as the regulatory relationships, signaling pathways, and the multi-protein machines that provide a mechanistic basis for a genes physiological effects. The conserved interactions and complexes described in this thesis are a step towards filling this gap. The shared code and methods enable others to apply this approach and build on it to map interactions across broad swathes of biology, uncovering machines and relationships that reveal novel mechanistic and physiological insights to increase our predictive understanding of the link from genes to phenotype and physiological effects.

Beyond these direct impact of undertaking this work and absorbing the biological knowledge in developing a useful lens into biology, I have learned a few important lessons I will carry forward with me in my ambitions to have the largest positive impact I can on the advancement of biological research towards understanding and treating all the limitations and failures of our amazing, but fragile, human bodies.

5.1 Large-scale, unbiased data

Biologists elucidated our current understanding of biology through low-throughput, rigorous investigations of diverse biological systems, often supported by the incredible power of human intuition. These investigations have historically focused predominantly on a small number of genes and proteins, a small enough set to fit in a simple cartoon diagram, and have employed a vast array of biochemical and molecular biology approaches to uncover the often highly complex relationships among them. Without such data gathered using these methods, the interpretation and validity of unbiased and high-throughput data sets would be difficult or impossible to determine. However, once such well-established knowledge exists, recent advances in experimental tools and methods in computational capacity and software are well-positioned to greatly expand its breadth and precision in biomedical research.

5.2 Functional basis for interrelatedness

The interrelatedness of different biological processes means that, in seeking to predict, for instance, protein complexes, you may consider using other types of biological relationships that at first glance appear unrelated. Nearly all aspects of biology relate to function, which is one way of wording the principle of guilt-by-association as applied to biology, indicating that one kind of biological relationship might be useful in predicting many other types of biological relationships (Wang 2010). This is especially true in the context of modern machine learning methods, which are excellent at distinguishing features that are predictive of a desired output from those that are not.

5.3 Orthogonal measurements in changing biological contexts

When the mutual information content between two data sets or two data sources is high, the amount of new understanding that can be uncovered is limited. For instance, identifying nearly 1,000 complexes based on co-fractionation experiments would be very challenging with a size-matched data set consisting only of experiments using the same fractionation method, the same cell type or tissue, the same species, and the same experimental conditions. Asking this same experimental question in highly varied biological contexts can dramatically increase our ability to extract meaningful biological insights.

5.4 Open-source computational tools in proteomics and network biology

The continued advances in genomics has convinced a growing portion of new and existing generations of biologists to the value of a systems perspective in biology. Biologists and labs integrating computational approaches to model and predict biology, especially at the genome level, continue to make rapid advancements in the field and benefit enormously from the plethora of tools available for everything from aligning genomes to interpreting genome-wide RNA-seq data sets. Due to the smaller relative size and newness of the field, proteomics and network biology lack the level of computational infrastructure available in genomics, which impedes progress to understand biological relationships and networks. It is my hope that my use of open-source tools whenever possible and publication of the methods and code for mass spectrometry data processing, data integration and inference, clustering, analysis and visualization will serve both to improve the suite of available open-source tools and methods and also to encourage others to develop and share open-source tools for these and related problems in the field.

5.5 Computational biologists at a point of high leverage

We are in the middle of an explosion of advancement in computational resources and techniques, and interest is high in computer science, software engineering, machine learning and artificial intelligence. While this explosion has undoubtedly led to steps forward in computational methods within biology, the vast majority of youths and adults with advanced computational skills never find themselves

working on or even considering problems relating to biology or medical research. In my opinion, this problem should be of utmost concern to those wishing to advance biological and medical research. This opinion has guided my career path into industry. I believe it may be possible to significantly increase the number and attractiveness of opportunities for computational scientists without a biology background to acquire one more easily, by providing avenues of entry into the field consisting of challenging and compelling problems that are approachable with minimal domain knowledge. As exciting problems and gifted co-workers are, in my view, one of the main drivers of career choices among the most ambitious and gifted computational workers, I hope this approach can contribute to the growth of a virtuous cycle, attracting more and more of the best and brightest to problems in the field.

5.6 The new way, not like the old way

Only 15 years ago, we saw the first draft of the human genome. The subsequent progress in biological understanding, if not yet medical treatments, has been phenomenal. Now we have single-cell genomes shedding light on the heterogeneity of tumors, and single-cell transcriptomes revealing the differentiation path of cellular networks. With further advances in co-fractionation and other approaches, we are on our way to establish personalized and, eventually, single-cell interaction networks. Our ability to generate rich biological data has grown tremendously and continues to increase rapidly. Many in the field complain today, as they have for decades, that we are drowning in a sea of biological data, far more of it than we even yet know how to properly analyze, much less fully exploit (Roos, 2001;

Brooksbank, Cameron, and Thornton, 2005; Marx, 2013). The mountain of mass spectrometry data generated for the studies described here is a perfect example. Our analyses only tapped a minuscule fraction of the information available in these experiments. If only we had more insight into the best questions to ask, more ideas for how to leverage this data to support of other lines of research, and more computational resources and talent to apply to the problem, I have little doubt the discoveries described here would be eclipsed many times over.

Biologists have started to see signs of the rising tide, but have yet to fully appreciate the enormity of the approaching wave that accompanies the coming computational revolution. A decade ago, a personal assistant on everyones phone capable of taking notes, sending emails, and scheduling meetings was a distant dream. Cars capable of driving themselves in nearly all conditions were widely perceived as science fiction or generations away, dreams of naive artificial intelligence and machine learning researchers and enthusiasts. And yet, rapid advances in machine learning have brought the first to fruition and the second to within a few years reach. I hope biologists will understand the connections between these and other advances in artificial intelligence and machine learning, as these connections foretell major changes coming to the field in the coming decade.

How do genomes, transcriptomes, protein interactions, genetic interactions, regulatory networks, chemogenomics, and other big data subfields of systems biology all fit together to drive progress on many fronts in biology and medical research? I do not know the answer, but I expect we will likely discover it in the coming decade. I predict that computational biologists with a combination of in-

tense scientific rigor and intimate familiarity with the descendants of today's cutting edge machine learning methods will lead the way.

Bibliography

Adamcsek, B et al. (2006). "CFinder: locating cliques and overlapping modules in biological networks". In: *Bioinformatics* 22.8, pp. 1021–1023.

Adams, Mark D. et al. (2000). "The Genome Sequence of Drosophila melanogaster". In: *Science* 287.5461, pp. 2185–2195.

Alberts, Bruce (1998). "The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists". In: *Cell* 92.3, pp. 291–294.

Bader, Gary D and Christopher WV Hogue (2003). "An automated method for finding molecular complexes in large protein interaction networks". In: *BMC Bioinformatcis* 4.1, p. 2.

Bandyopadhyay, S (2006). "Systematic identification of functional orthologs based on protein network comparison". In: *Genome Research* 16.3, pp. 428–435.

Bandyopadhyay, Sourav et al. (2008). "Functional Maps of Protein Complexes from Quantitative Genetic Interaction Data". In: *PLoS Computational Biology* 4.4, pp. 1–8.

Beck, Martin et al. (2011). "The quantitative proteome of a human cell line". In: *Molecular Systems Biology* 7, pp. 1–8.

Beeckmans, Sonia (1999). "Chromatographic Methods to Study Protein-Protein Interactions". In: *Methods* 19.2, pp. 278–305.

Bezginov, Alexandr et al. (2013). "Coevolution Reveals a Network of Human Proteins Originating with Multicellularity". In: *Molecular Biology and Evolution* 30.2, pp. 332–346.

Brent, R. and M.S. Ptashne (1989). *Regulation of eucaryotic gene expression*. US Patent 4,833,080. URL: `https://www.google.com/patents/US4833080`.

Brohée, Sylvain and Jacques van Helden (2006). "Evaluation of clustering algorithms for protein-protein interaction networks." In: *BMC Bioinformatcis* 7, p. 488.

Brooksbank, Catherine, Graham Cameron, and Janet Thornton (2005). "The European Bioinformatics Institute's data resources: towards systems biology". In: *Nucleic Acids Research* 33.suppl 1, pp. D46–D53.

Burstein, Ezra et al. (2005). "COMMD proteins, a novel family of structural and functional homologs of MURR1." In: *J. Biol. Chem* 280.23, pp. 22222–22232.

Cox, Jürgen and Matthias Mann (2008). "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification". In: *Nature Biotechnology* 26.12, pp. 1367–1372.

Croft, David et al. (2014). "The Reactome pathway knowledgebase". In: *Nucleic Acids Research* 42.D1, pp. D472–D477.

Dalquen, Daniel A. et al. (2013). "The Impact of Gene Duplication, Insertion, Deletion, Lateral Gene Transfer and Sequencing Error on Orthology Inference: A Simulation Study". In: *PLoS ONE* 8.2, pp. 1–11.

Diament, Benjamin J. and William Stafford Noble (2011). "Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra". In: *Journal of Proteome Research* 10.9, pp. 3871–3879.

Eisenberg, Eli and Erez Y. Levanon (2003). "Preferential Attachment in the Protein Network Evolution". In: *Phys. Rev. Lett.* 91 (13), p. 138701.

Elias, Joshua E and Steven P Gygi (2007). "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry". In: *Nature Methods* 4.3, pp. 207–214.

Eng, Jimmy K., Ashley L. McCormack, and John R. Yates (1994). "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database". In: *Journal of the American Society for Mass Spectrometry* 5.11, pp. 976–989.

Fields, S and O Song (1989). "A novel genetic system to detect protein protein interactions". In: *Nature* 340.6230, pp. 245–246.

Finn, Robert D. et al. (2014). "Pfam: the protein families database". In: *Nucleic Acids Research* 42.D1, pp. D222–D230.

Fraser, Hunter B. and Joshua B. Plotkin (2007). "Using protein complexes to predict phenotypic effects of gene mutation". In: *Genome Biology* 8.11, pp. 1–9.

Gandhi, T K B et al. (2006). "Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets". In: *Nature Genetics* 38.3, pp. 285–293.

Gavin, Anne-Claude et al. (2006). "Proteome survey reveals modularity of the yeast cell machinery". In: *Nature Publishing Group* 440.7084, pp. 631–636.

Goffeau, A. et al. (1996). "Life with 6000 Genes". In: *Science* 274.5287, pp. 546–567.

Güldener, Ulrich et al. (2006). "MPact: the MIPS protein interaction resource on yeast". In: *Nucleic Acids Research* 34.suppl 1, pp. D436–D441.

Guruharsha, K. G. et al. (2011). "A Protein Complex Network of Drosophila melanogaster". In: *Cell* 147.53, pp. 690–703.

Hart, G Traver, Insuk Lee, and Edward M Marcotte (2007). "A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality". In: *BMC Bioinformatcis* 8.1, p. 236.

Hart, G. Traver, Arun K. Ramani, and Edward M. Marcotte (2006). "How complete are current yeast and human protein-interaction networks?" In: *Genome Biology* 7.11, pp. 1–9.

Hartwell, Leland H et al. (1999). "From molecular to modular cell biology". In: *Nature* 402.6761 Suppl, pp. C47–C52.

Havugimana, Pierre C. et al. (2012). "A Census of Human Soluble Protein Complexes". In: *Cell* 150.5, pp. 1068–1081.

Hein, Marco Y et al. (2015). "A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances". In: *Cell* 163.3, pp. 712–723.

Hu, Pingzhao et al. (2009). "Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins." In: *PLoS Biology* 7.4, e96.

Hunt, Donald F. et al. (1986). "Protein sequencing by tandem mass spectrometry". In: *Proceedings of the National Academy of Sciences* 83.17, pp. 6233–6237.

Huttlin, Edward L., Lily Ting, et al. (2015). "The BioPlex Network: A Systematic Exploration of the Human Interactome". In: *Cell* 162.2, pp. 425–440.

Huttlin, Edward L., L. Ting, et al. (2016). *High-Throughput Proteomic Mapping of Human Interaction Networks via Affinity-Purification Mass Spectrometry*. Pre-Publication Dataset. URL: `http://thebiogrid.org/166968/publication/`.

Jansen, Ronald et al. (2003). "A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data". In: *Science* 302.5644, pp. 449–453.

Jeong, H et al. (2001). "Lethality and centrality in protein networks". In: *Nature* 411.6833, pp. 41–42.

Kanehisa, Minoru and Susumu Goto (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1, pp. 27–30.

Kelley, Brian P. et al. (2003). "Conserved pathways within bacteria and yeast as revealed by global protein network alignment". In: *Proceedings of the National Academy of Sciences* 100.20, pp. 11394–11399.

Kim, Min-Sik et al. (2014). “A draft map of the human proteome”. In: *Nature* 509.7502, pp. 575–581.

Kirkwood, Kathryn and Angus I Lamond (2013). “Characterisation of Native Protein Complexes and Protein Isoform Variation using Size-Fractionation Based Quantitative Proteomics”. In: pp. 1–55.

Knoll, Andrew H. (1992). “The early evolution of eukaryotes: a geological perspective”. In: *Science* 256.5057, pp. 622–627.

Kolanczyk, Mateusz et al. (2015). “Missense variant in CCDC22 causes X-linked recessive intellectual disability with features of Ritscher-Schinzel/3C syndrome”. In: *European Journal of Human Genetics* 23.5, pp. 633–638.

Krey, Jocelyn F et al. (2014). “Accurate Label-Free Protein Quantitation with High- and Low-Resolution Mass Spectrometers”. In: *Journal of Proteome Research* 13.2, pp. 1034–1044.

Krogan, Nevan J et al. (2006). “Global landscape of protein complexes in the yeast Saccharomyces cerevisiae”. In: *Nature Publishing Group* 440.7084, pp. 637–643.

Kwon, Taejoon et al. (2011). “MSblender: A Probabilistic Approach for Integrating Peptide Identifications from Multiple Database Search Engines”. In: *Journal of Proteome Research* 10.7, pp. 2949–2958.

Lage, Kasper et al. (2007). “A human phenome-interactome network of protein complexes implicated in genetic disorders”. In: *Nature Biotechnology* 25.3, pp. 309–316.

Lee, Insuk et al. (2011). “Prioritizing candidate disease genes by network-based boosting of genome-wide association data”. In: *Genome Research* 21.7, pp. 1109–1121.

Link, A J et al. (1999). “Direct analysis of protein complexes using mass spectrometry”. In: *Nature Biotechnology* 17.7, pp. 676–682.

Lynch, Iseult et al. (2007). “The nanoparticle-protein complex as a biological entity; a complex fluids and surface science challenge for the 21st century”. In: *Advances in Colloid and Interface Science* 134–135. Surface forces: wetting phenomena, membrane separation, rheology. Topical issue in honour of Victor Starov, pp. 167–174.

Malovannaya, Anna et al. (2011). “Analysis of the Human Endogenous Coregulator Complexome”. In: *Cell* 145.5, pp. 787–799.

Marx, Vivien (2013). “Biology: The big challenges of big data”. In: *Nature* 498.7453, pp. 255–260.

McDonald, Fiona J. (2013). “COMMD1 and ion transport proteins: what is the COMMection? Focus on ‘COMMD1 interacts with the COOH terminus of NKCC1 in Calu-3 airway epithelial cells to modulate NKCC1 ubiquitination’”. In: *American Journal of Physiology - Cell Physiology* 305.2, pp. C129–C130.

Nakhleh, Luay, Donald A. Ringe, and Tandy Warnow (2005). “Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages”. In: *Language* 81.2, pp. 382–420.

Nepusz, Tamas, Haiyuan Yu, and Alberto Paccanaro (2012). “Detecting overlapping protein complexes in protein-protein interaction networks”. In: *Nature Methods* 9.5, pp. 471–472.

O’Farrell, Patrick H. (1975). “High resolution two-dimensional electrophoresis of proteins”. In: *Journal of Biological Chemistry* 250.10, pp. 4007–4021.

Oliphant, Travis E (2007). “Python for Scientific Computing”. In: *Computing in Science & Engineering* 07, pp. 10–20.

Oliver, Stephen G (2000). “Proteomics: guilt-by-association goes global”. In: *Nature* 403.6770, pp. 601–603.

Ovádi, Judit (1995). *Cell Architecture and Metabolite Channeling*. Molecular Biology Intelligence Unit. R.G. Landes Co.

Pedregosa, Fabian et al. (2011). “Scikit-learn: Machine learning in Python”. In: *The Journal of Machine Learning Research* 12, pp. 2825–2830.

Remm, Maido, Christian EV Storm, and Erik LL Sonnhammer (2001). “Automatic clustering of orthologs and in-paralogs from pairwise species comparisons”. In: *Journal of Molecular Biology* 314.5, pp. 1041–1052.

Rolland, Thomas et al. (2014). “A Proteome-Scale Map of the Human Interactome Network”. In: *Cell* 159.5, pp. 1212–1226.

Roos, David S. (2001). “Bioinformatics—Trying to Swim in a Sea of Data”. In: *Science* 291.5507, pp. 1260–1261.

Rual, Jean-François et al. (2005). “Towards a proteome-scale map of the human protein–protein interaction network”. In: *Nature* 437.7062, pp. 1173–1178.

Rubin, Gerald M. et al. (2000). “Comparative Genomics of the Eukaryotes”. In: *Science* 287.5461, pp. 2204–2215.

Ruepp, Andreas et al. (2008). “CORUM: the comprehensive resource of mammalian protein complexes”. In: *Nucleic Acids Research* 36.suppl 1, pp. D646–D650.

Sharan, R et al. (2005). “Conserved patterns of protein interaction in multiple species”. In: *Proceedings of the National Academy of Sciences* 102.6, pp. 1974–1979.

Singh, Rohit, Jinbo Xu, and Bonnie Berger (2008). “Global alignment of multiple protein interaction networks with application to functional orthology detection”. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.35, pp. 12763–12768.

Sonnhammer, Erik L.L. et al. (2014). “Big data and other challenges in the quest for orthologs”. In: *Bioinformatics* 30.21, pp. 2993–2998.

Stark, Chris et al. (2006). “BioGRID: a general repository for interaction datasets”. In: *Nucleic Acids Research* 34.suppl 1, pp. D535–D539.

Stelzl, Ulrich et al. (2005). “A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome”. In: 122.6, pp. 957–968.

Stumpf, Michael P. H. et al. (2008). “Estimating the Size of the Human Interactome”. In: *Proceedings of the National Academy of Sciences* 105.19, pp. 6959–6964.

The ENCODE Project Consortium (2012). “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414, pp. 57–74.

The Human Genome Project (2001). “Initial Sequencing and Analysis of the Human Genome”. In: *Nature* 409.6822, pp. 860–921.

Thiele, Ines et al. (2013). “A Community-Driven Global Reconstruction of Human Metabolism”. In: *Nature Biotechnology* 31.5, pp. 419–425.

Tillier, Elisabeth R.M. and Robert L. Charlebois (2009). “The Human Protein Coevolution Network”. In: *Genome Research* 19.10, pp. 1861–1871.

Turner, Brian et al. (2010). “iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence.” In: *Database : the journal of biological databases and curation* 2010, baq023.

Typas, Athanasios and Victor Sourjik (2015). “Bacterial protein networks: properties and functions”. In: *Nature Publishing Group* 13.9, pp. 559–572.

Uhlén, Mathias, Linn Fagerberg, et al. (2015). “Tissue-based map of the human proteome”. In: *Science* 347.6220.

Uhlén, Mathias, Per Oksvold, et al. (2010). “Towards a knowledge-based Human Protein Atlas”. In: *Nature Biotechnology* 28.12, pp. 1248–1250.

van de Sluis, Bart et al. (2002). “Identification of a new copper metabolism gene by positional cloning in a purebred dog population”. In: *Human Molecular Genetics* 11.2, pp. 165–173.

van Dongen, Stijn Marinus (2000). “Graph clustering by flow simulation”. PhD thesis. University of Utrecht, pp. 1–173.

Vidal, Marc and Stanley Fields (2014). “The yeast two-hybrid assay: still finding connections after 25 years”. In: *Nature Publishing Group* 11.12, pp. 1203–1206.

Vinayagam, A et al. (2013). “Protein Complex-Based Analysis Framework for High-Throughput Data Sets”. In: *Science Signaling* 6.264, rs5–rs5.

Voineagu, I et al. (2012). “CCDC22: a novel candidate gene for syndromic X-linked intellectual disability”. In: *Molecular Psychiatry* 17.1, pp. 4–7.

Von Mering, Christian et al. (2002). “Comparative assessment of large-scale data sets of protein-protein interactions”. In: *Nature* 417.6887, pp. 399–403.

Wan, Cuihong, Blake Borgeson, et al. (2015). “Panorama of Ancient Metazoan Macromolecular Complexes”. In: *Nature* 525.7569, pp. 339–344.

Wan, Cuihong, Jian Liu, et al. (2013). “ComplexQuant: High-throughput computational pipeline for the global quantitative analysis of endogenous soluble protein complexes using high resolution protein HPLC and precision label-free LC/MS/MS”. In: *Journal of Proteomics* 81. Special Issue: From protein structures to clinical applications, pp. 102–111.

Wang, Peggy I and Edward M Marcotte (2010). “It’s the machine that matters: Predicting gene function and phenotype from protein networks”. In: *Journal of Proteomics* 73.11, pp. 2277–2289.

Wilhelm, Mathias et al. (2014). “Mass-spectrometry-based draft of the human proteome.” In: *Nature* 509.7502, pp. 582–587.

Yanai, Itai et al. (2011). “Mapping Gene Expression in Two ¡em¿Xenopus¡/em¿ Species: Evolutionary Constraints and Developmental Flexibility”. In: *Developmental Cell* 20.4, pp. 483–496.

Zhang, Qiangfeng Cliff et al. (2012). “Structure-based prediction of protein-protein interactions on a genome-wide scale”. In: *Nature*, pp. 1–6.

Vita

Blake Borgeson was born (according to his parents and county records) in Austin, Texas, the city where he grew up, made and lost friends, played soccer and basketball, learned to swim, and still feels most at home. He went to Rice University with the intent to study physics, but found the math of information more compelling than the math of physical entities and switched to electrical engineering. After graduating, he spent a year working in Switzerland on account of the Alps and the mix of languages, then decided to try to become internet entrepreneur. Struggles and success emerged back in Austin, in the form of the online custom printing company BuildASign.com. Soon Blake found himself again following the call of science and pursued machine learning and systems biology in the Marcotte lab. Discovering with joy that he could in fact all at once be an entrepreneur and a scientist and enjoy the mountains, he co-founded Recursion Pharmaceuticals in Utah, where he hopes to advance his personal agendas of accelerating the end of disease and helping to ensure the survival and thriving of the human race.

This dissertation was typed by the author.

Permanent address: Reach me on the web at `http://blakeb.org`